AUTOMATICALLY IDENTIFYING FACET ROLES FROM COMPARATIVE STRUCTURES TO SUPPORT BIOMEDICAL TEXT SUMMARIZATION

BY

ANA LUČIĆ

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library & Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Associate Professor Catherine Blake, Chair and Research Director
Associate Professor Roxanna Corina Girju
Associate Professor Miles Efron
Professor Allen Renear
Professor J. Stephen Downie

ABSTRACT

Within the context of biomedical scholarly articles, comparison sentences represent a rhetorical structure commonly used to communicate findings. More generally, comparison sentences are rich with information about how the properties of one or more entities relate one another. So far, in the biomedical domain, the emphasis has been on recognizing comparative sentences in the text. This dissertation goes beyond sentence-level recognition and aims to automate the identification of the integral parts of a comparison sentence which are called *comparative facets* and include: compared entities, the basis or the endpoint of comparison as well as the result or the relationship that binds the entities and the basis. Only the sentences that contain each of the four facets are of interest in this thesis.

With respect to the first compared entity, the system achieves an average F_1 on a random sample of short (between 11 and 21 words long) sentences of 0.65; medium (between 22 and <= 28 words) sentences 0.70; long (between 29 and <=36 words) sentences 0.60 and very long (more than 36 words), 0.60. With respect to the basis of comparison prediction (the endpoint), the average F_1 measure ranged from 0.66 on short, 0.57 on medium, 0.56 on long, and 0.50 on very long sentences. The average F_1 achieved with respect to the second entity compared ranged from 0.91 on short, 0.85 on medium, 0.81 on long and 0.72 on very long sentences. In the area of semantic relation identification, the performance achieved was also sensitive to sentence length: the average F_1 measure on short sentences was 0.80; it was 0.71, 0.56, and 0.51 on medium, long, and very long sentences respectively. Thus, the methods developed in this dissertation work better on sentences that are shorter (<= 28 words) and on those that do not contain multiple claims or disjunctive conjunctions. When applied to a previously unseen collection of *breast cancer* articles, the performance achieved with respect to the identification of compared entities and the endpoint was comparable to the results achieved on the collection that was used for building and testing the models. This result is promising with respect to the potential of this model being applied on other collections of scholarly articles in the biomedical sciences.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the support from my family. Special thanks goes out to my husband, George Gasyna, who, more than ten years ago, had gone through the dissertation writing process himself and perfectly understands the ills and pains associated with it. I am most grateful to our two daughters, Evie and Sophie, for their patience and understanding during the times I had to be away from them. I am also most grateful to my sister, Nataša Milosavljević and my parents, Milanka and Ratko Lučić for their encouragement and support during the lengthy process of dissertation writing. I am very grateful for my colleagues from the Socio-Technical Data Analytics (SODA) research group, Henry Gabb, Jooho Lee, and Jinlong Guo—also known as Kenney—for sharing their code, expertise, and their encouragement and support along this way. The underlying data and preprocessing of it would not have been possible without the expertise of Craig Evans who, on numerous occasions, preprocessed data and generously shared his code and programming knowledge with the rest of the research group. I am very grateful to Dr. Rhiannon Bettivia for reading my presentations, papers, and applications and suggesting improvements. I am deeply indebted to one of my professor from the University of Belgrade, later on also a friend, Dr. Ileana Čura, for inspiring my passion for knowledge and learning.

I would like to thank my committee members, Dr. Roxanna Corina Girju, Dr. Allen Renear, Dr. Miles Efron, and Dr. Stephen J. Downie for their guidance, feedback, ideas, and discussions throughout these past five years. I also thank them for challenging me to see things from a different perspective. I am very grateful for the wonderful faculty members and staff of the School of Information Sciences. In particular, I am grateful for Vetle Torvik's Data Mining class that is, at least partially, responsible for Text Mining becoming one of my areas of research interest. Finally, I would like to extend my gratitude and deep appreciation to my advisor and Dissertation Chair, Dr. Catherine Blake. In addition to being a stellar researcher Dr. Blake is also a wonderful mentor and a role model whose creativity never ceases to amaze me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Comparison sentences represent a rhetorical structure that frequently, although not exclusively, express how the properties of one entity relate to those of a compared entity. In addition to being commonly used to report the results of empirical research, comparisons have also been identified as a common information need. An analysis of clinical questions in the National Library of Health (NLH) Question Answering Service (http://www.clinicalanswers.nhs.uk) revealed that 16% of the 4,580 questions referred to direct comparisons of different drugs, treatment methods and intervention (Leonhard, 2009). And yet, the current information systems do not provide a method of answering comparative queries that deal with how one entity has been compared to another with respect to the properties they were compared on and with respect to the results achieved. The goal of this dissertation is to track and eventually aggregate the information on how one entity or group of entities has been compared to another through a collection of articles. For example, if the user was interested in discovering ways in which one anti-diabetic drug has been compared to another anti-diabetic drug they might use the PubMed search engine to find this information. Data about these two drugs is available and yet the format through which we can zoom in specifically on the information that relates how the two drugs compare is not within an easy reach. We can modify the query and filter the search results to only human clinical trial which would reduce the number of results, and yet, sifting through any number of clinical trials in search of a particular information is not an easy task. The goal of this dissertation is to extract the prominent information from comparative sentences in scholarly articles such as the drugs, treatments, interventions of interest, the basis on which the drugs were compared, and also the result of the comparison. These extracted facets—as they will be called throughout of the rest of this document—can then be used to populate a comparative summary. The comparative summary that, at this stage, is only envisioned to be populated with extracted facets from comparative sentences belongs to a type of an extractive summary that is particularly appropriate in medical domain where generative summary approaches are not sufficient (Nenkova, 2007). Additionally, such a summary represents a type of strategic reading of scholarly literature, the one that allows “rapid and productive engagement with the literature” (Renear & Palmer, 2009) that would be of interest to general users but also professionals and decision makers who need quick and efficient summaries of information. More generally, the structure of comparison sentences and the methods that allow parsing of comparison structure in an automated can be seen as particularly relevant to Comparative Effectiveness Research whose goal is to provide evidence on the effectiveness, benefits and drawbacks of different treatment options. The evidence is generated from studies that compare drugs, medical devices, tests, surgeries, or ways to deliver health care. Identifying and extracting comparison facets in an automated way highlight the areas where comparative work has or has not been done.

Identifying comparison sentences as a way of summarizing the information expressed through comparisons has been a focus of several studies so far (Jindal & Liu, 2006; Fiszman et al., 2007; Hoon Park & Blake, 2012). The main contribution of this dissertation is that it goes a level deeper than the sentence level: it identifies the *integral parts of a comparison sentence*—the previously mentioned *facets*. To define the integral parts of comparison sentences this study relies on the Claim Framework (CF) model in the background. The Claim Framework specifies five types of claims scientists commonly use to communicate the findings of their experiments: explicit, implicit, observations, correlations, and comparisons. Each of these types of claims can be expressed through four facets: entity 1, entity 2, endpoint of comparison, and nature of change (Blake, 2010). In this study, only those sentences that contain each of the four facets are included in the analysis. In general, the nature of change corresponds to the relationship that binds the two entities (entity 1 and entity 2) and the endpoint of comparison. Relying on the CF model in the background, this dissertation develops a method that can tease out the integral parts of comparison sentences that are envisioned as playing a role of content elements in a comparative summary.

1.1 STATEMENT OF THE PROBLEM

Although not a frequent rhetorical structure comparison sentences are a fairly complex rhetorical structure and have been called “almost notorious for its syntactic complexity” but also for their semantic complexity which constitutes the main reason why they are difficult to process automatically (Bresnan, 1973; Friedman, 1989). Put differently, numerous ways of expressing comparisons complicate the task of automatically processing a comparison sentence and identifying their integral parts.

Jindal & Liu (2006) proposed an enhanced categorization of comparative sentences that divides comparative sentences into gradable and non-gradable. Three types of gradable sentences are identified: non-equal gradable, equative, and superlative. Gradable comparisons express an ordering of entities with respect to an aspect whereas non-gradable comparisons compare features of two or more entities but do not grade them. Of the three gradable types proposed by Jindal & Liu (2006), two are the focus of this thesis: non-equal gradable and equative. An example of a non-equal gradable sentence is the following:

> HbA_1c was *higher* in diabetic women than men (P = 0.004). $_{12031985}$[1]

In this sentence, diabetic women and men were compared with respect to HbA_1c (glycated haemoglobin) and it was established that the level of glycated haemoglobin was higher in diabetic women than men. The adjective higher indicates that one group or subject had a *higher* level of HbA_1c than the other and hence they are considered non-equal comparative sentences. Superlative sentences, however, are not the subject of this study and the section 1.2 will provide an explanation about why this group is excluded from consideration.

Non-gradable comparisons express how two entities or concepts are *similar* or *different* on one or more aspects. For example, consider the following sentence:

> ADX + CORT animals had plasma corticosterone concentrations *similar* to sham animals (6.3 2.8 microg/dl) (P > 0.05). $_{14633853}$[2]

In this example, ADX + CORT animals are *similar* to sham animals with respect to plasma corticosterone concentrations and it is not the case that one type of animal had more or less of plasma corticosterone concentrations, their level of plasma corticosterone concentrations were similar.

In general, both gradable and non-gradable expressions of comparisons are covered by the methods proposed in this thesis. The division of comparison sentences into gradable and non-gradable, however, is not the only way comparison sentences can be categorized. Comparative

[1] This sentence is annotated and discussed later in the dissertation (sentence 56)
[2] This sentence is annotated and discussed later in the dissertation (sentence 46)

sentences not only differ with respect to the type of comparative relation they convey but also with respect to the amount of information they convey. For example, a sentence sometimes merely expresses that a comparison was made between the two entities without providing the details about the aspect or endpoint on which the comparison was made. Sentences 1 and 2 are a good example:

(1) We compared control [entity 1] with treated animals [entity 2]. 12538615

(2) Parous rats [entity 1] were compared with respective age-matched virgins (AMVs) [entity 2]. 10223190

These two sentences only indicate a statement that a comparison was made, they do not indicate the result of comparison or the aspect/basis on which the entities were compared. Sometimes, however, a comparison sentence expresses that two entities or concepts were compared on a certain aspect or *endpoint*—as this concept will be referred to throughout the rest of the document—but they do not include the information about the result of the comparison such as in sentences 3 and 4:

(3) We compared arterial pressure [endpoint] between mice [entity 1] on normal [entity 1] or high-fat diet [entity 2]. 15983201

(4) This study compared the frequency of p53 mutations [endpoint] in BRCA1-associated breast carcinomas [entity 1] with that in sporadic breast tumors [entity 2] in a prevalence sample of Ashkenazi Jewish women [entity 2]. 10070948

Yet the third type of a comparison sentence includes all of this information: the entities that were compared, the endpoint and also the result as in sentence 5:

(5) The result showed that ethanol feeding [entity 1] in rats [entity 1] *increased* [relation] c-Jun mRNA level [endpoint], as compared with the non-ethanol-fed rats [entity 2]. 11470752

Comparison sentences do not only contain a varying amount of information but can also include more than one individual claim, such as in sentence 6:

(6) The plasma insulin concentration [endpoint_A] at 8 weeks [endpoint_A] of age [endpoint _A] and the pancreatic insulin content [endpoint_B] and the beta-cell mass [endpoint_C] on day 8 [endpoint_BC] and 8 weeks [endpoint_BC] of age [endpoint_BC] in STZ-treated rats [entity 1] *were severely* [relation modifier] *reduced* [relation] compared with those of normal rats [entity

2] (P < 0.001). 14988244

More particularly, sentence 6 contains three individual claims that can be expressed as:

I. The plasma insulin concentrations [endpoint_A] at 8 weeks of age [was] *severely* [relation modifier] *reduced* [relation] in STZ-treated rats [entity 1] compared with those of normal rats [entity 2].

II. The pancreatic insulin content [endpoint_B] at 8 weeks of age [was] *severely* [relation modifier] *reduced* [relation] in STZ-treated rats [entity 1] compared with those of normal rats [entity 2].

III. The beta-cell mass [endpoint_C] at 8 weeks of age [was] *severely* [relation modifier] *reduced* [relation] in STZ-treated rats [entity 1] compared with those of normal rats [entity 2].

In addition to multiple endpoints, a comparison sentence can compare multiple entities, all within the context of one sentence, such as in sentence 7:

(7) Compared with non- [entity 2_A] and irregular tea drinkers [entity 2_B] in particular, we found *reduction* [relation_A] in circulating levels [endpoint_AB] of both estrone (-13%) [endpoint_A] and estradiol (- 8%) [endpoint_B] among weekly/daily green tea drinkers [entity 1_A] and *increase* [relation_B] in both estrone (+19%) [endpoint_C] and estradiol (+10%) levels [endpoint_D] among weekly/daily black tea drinkers [entity 1_B]. 15661801

This sentence contains eight claims each of which contains four facets that are integral to a comparison claim:

I. Weekly/daily green tea drinkers [entity 1_A] had *reduced* [relation_A] circulating levels of estrone (13%) [endpoint_A] compared with non- tea drinkers [entity 2_A].

II. Weekly/daily green tea drinkers [entity 1_A] had *reduced* [relation_A] circulating levels of estradiol (-8%) [endpoint_B] compared with non- tea drinkers [entity 2_A].

III. Weekly/daily black tea drinkers [entity 1_B] had *increased* [relation_B] circulating levels of estrone (+19%) [endpoint_C] compared with non- tea drinkers [entity 2_A].

IV. Weekly/daily black tea drinkers [entity 1_B] had *increased* [relation_B] circulating levels of estradiol (+10%) [endpoint_D] compared with non- tea drinkers [entity 2_A].

V. Weekly/daily green tea drinkers [entity 1_A] had *reduced* [relation_A] circulating levels of estrone (-13%) [endpoint_A] compared with irregular tea drinkers [entity 2_B].

VI. Weekly/daily green tea drinkers [entity 1_A] had *reduced* [relation_A] circulating levels of estradiol (-8%) [endpoint_B] compared with irregular tea drinkers [entity 2_B].

VII. Weekly/daily black tea drinkers [entity 1_B] had *increased* [relation_B] circulating levels of estrone (+19%) [endpoint_C] compared with irregular tea drinkers [entity 2_B].

VIII. Weekly/daily black tea drinkers [entity 1_B] had *increased* [relation_B] circulating levels of estradiol (+10%) [endpoint_D] compared with irregular tea drinkers [entity 2_B].

As we see from this example, as the number of entities and endpoints in the sentence increases so does the number of individual claims. Comparison sentences not only differ with respect to the amount of information they contain but also with respect to the number of individual claims they express. This variety in the structure (syntax), type (gradable and non-gradable), and the amount of information they convey naturally have ramifications on the ability to extract the information from comparison sentences in an automated way. And yet the ability to identify individual claims as well as their integral parts can bring us closer to summarizing and synthesizing the information that has been shared through comparative structures within a single or even or even across domains.

Given the inherent complexity of comparison sentences and multiple ways of expressing them, the methods proposed in this study do not attempt to cover all the ways in which comparisons can be expressed. The ways of expressing comparison relations are diverse and manifold and it is difficult, almost impossible, to imagine a system that would subsume all expressions of comparisons. The methods for doing this would, most likely, need to be diverse and manifold, just like comparison sentences themselves. What the methods in this dissertation do propose, however, is a way of extracting most relevant facets from a *particular kind* of a comparison sentence. This particular kind of comparison sentence from now on will be referred to as *the direct comparison*: the comparative sentence that juxtaposes two or more entities, compares them on one or more properties and also reports the results of the comparison that has occurred.

The following sub-section further defines direct comparison by specifying other types of comparisons that are not of interest in this study and demonstrating how the particular type that is of interest in this study is different from other types.

1.2 DIRECT COMPARISON VERSUS OTHER TYPES OF COMPARISONS

To better understand the nature and characteristics of direct comparative sentences and to differentiate this type from other types, the following outline highlights comparative sentences that are *not* considered in this thesis.

1.2.1 Indirect comparison sentences

Although a nice complement to direct comparison sentences, indirect comparisons are not considered in this study. A type of comparative relation of interest in this dissertation is the one that directly compares two entities. Not only does it directly compare entities but it also reports the basis on which the entities were compared as well as the result or the outcome of the comparison. A head-to-head comparison of two interventions and their result all communicated within the confines of a single sentence constitutes a *direct* comparison sentence. And yet, high quality evidence consisting of systematic review of randomized clinical trials that provide direct (head-to-head) comparison of two interventions are commonly rare, sometimes non-existent or inconclusive. This is why recently there has been a shift towards the identification of indirect comparisons in scholarly articles (Donegan et al., 2010; Guichard, et al., 2015). Occasionally, indirect comparisons can be more reliable than direct evidence due to methodological inadequacies of trials (Song et al., 2008). This dissertation focuses on direct evidence which, although infrequent, is packed with valuable information that this work aims to unpack with the aim of viewing the result of such evidence in aggregate.

1.2.2 Simile

A figure of speech that is related to comparisons and through which two unlike things are compared to each other, *simile*, is not considered in this study. Computational approaches to simile have been explored (Niculae & Yaneva, 2013) with respect to a decreased ability of autistic people to understand metaphors and figurative language. Although not very common in the biomedical scholarly literature, simile and metaphors are rather common in literary works. Frequently, although not exclusively, simile use the preposition *like* to express the two unlike entities that do not have to be of the same kind. And yet, the preposition *like* does not necessarily indicate a simile: sometimes, the preposition *like* indicates a comparison between the same type of entities, as in the following example:

(8) Anatabine [entity 1_A] and anabasine [entity 1_B] like nicotine [entity 2_A] and cotinine [entity_2_B] are non-carcinogenic. 12082012

In this example, the tobacco alkaloids *anatabine*, *anabasine*, *nicotine* and *cotinine* all belong to the same type of entity, *alkaloids*, and the preposition *like* is used to compare them rather than to exaggerate the characteristics of any. Although comparisons that contain the preposition *like* but are not simile occur relatively frequently in the biomedical scholarly literature, this type of comparison is not a focus of this study. The comparative relation that is expressed using the preposition *like* differs from the relations that are expressed using the comparison anchor such as, for example, *compared with*. In sentence 8, "non-carcinogenic" is not the result of a comparison of *anatabine* and *anabasine,* it represents a joint property of both pairs of entities. Direct comparisons, on the other hand, indicate a change that has occurred or not occurred as a result of comparing two or more entities.

1.2.3 Deviations from Claim Framework

Among the types of comparisons that are not the focus of this study are those in which the information conveyed does not follow the Claim Framework model (sentence 9):

(9) Hence, it is possible that the TGF-beta1 growth response is more dependent on the amount of TbetaR-II receptor expression than it is on the TGF-beta1-PRL response. 9681516

Sentence 9 communicates that TGF-beta1 growth response was compared against two receptor expressions: *TbetaR-II receptor expression* and *TGF-beta1-PRL response*. Although certainly a comparison, the relationship expressed in this sentence is different from the relationship that expresses that two entities were compared on a particular aspect. In this sentence, one entity is compared against other two entities which represents a different relation from the one in which two entities are compared against a common property.

1.2.4 Superlative

This dissertation also does not take into consideration superlative relation because superlative represents a different type of relation. Mainly, superlatives express how one entity compares to a set of entities. Given that it is difficult to infer using automated methods the entities that comprise a set that an entity was compared to this makes superlatives even harder to process automatically than comparative sentences (Sheible, 2007).

1.2.5 Lack of one or more of the integral components (facets)

Sentences that do not contain one of the required comparison facets for a direct comparison (the minimum of two compared entities, the endpoint and the result of comparison or the main predicate or the relation) are not considered in this study. Even if the component or facet of a comparison sentence that is of interest in this work is omitted but can be inferred from the surrounding context (anaphoric reference), this sentence is not used for either training or testing the model. Consider sentence 10:

(10) This is 262 times lower than the median levels found in Chinese workers. 15817613

Although sentence 10 contains a comparative adjective and a comparison anchor (*lower than*), it is not clear what level is being compared. Although it may be possible to infer the context of comparison from the preceding sentences, this is left for future work.

Chapter 3 of this dissertation contains a more detailed analysis and the categorization of the additional types of comparisons that are found in scholarly articles that are also not considered direct comparative sentences. For example, comparison sentences that are speculative in nature, those that are directly connected to the method of conducting a study rather than the result of the study and the cases in which the comparison anchor which, most commonly, is an indicator of a comparative relation expresses an explicit rather than a comparative statement.

This thesis focuses on sentences that include each of the following facets of a comparison claim: (at least) two compared entities, endpoint or the basis on which the entities were compared as well the relationship that binds the three facets. Thus, this dissertation distills one particular type of comparison sentence that although not very frequent is yet common and filled with information that this work strives to make available for knowledge representation and further computational analysis purposes. As Chapter 5 of this dissertation will demonstrate, this particular type of a comparison subsumes a number of comparative predicates and as such, can be considered a type of a *meta-relation*, a higher and more abstract type of relation.

1.3 LITERATURE BACKGROUND

Communication through journals and conference papers, if studied in aggregate or through a corpus of articles rather than on individual basis, has the capability of revealing certain trends and patterns. In 1988, Zellig Harris argued that language in sciences is limited in the sense that there are more restrictions on which words can co-occur and which words can entail other words. These restrictions are the reason why it is easier to translate scientific than literary texts. Through his study of immunology literature that spanned the years 1935-66 Harris showed how the discussion and the controversy about which cell produced the antibodies lymphocyte or plasma cells was resolved in the field of immunology (1988). By converting the statements in the articles into their formulaic expressions, Harris revealed that it was possible to track the development of this particular controversy through a corpus of articles and eventually the resolution of the controversy which came in the form of the following statement: "Lymphocytes develop into plasma cells" (Harris, 1988). Thirty years later, it seems as though the advances in Natural Language Processing area have made it possible to identify and create the formulaic sublanguage that Harris anticipated in his research. The existence of software that can recognize named entities of different types in an automatic way, such as the Stanford Named Entity Recognizer (nlp.stanford.edu/software/CRF-NER.shtml) (Finkel et al., 2005) or Genia tagger (nactem.ac.uk/tsujii/GENIA/tagger/) (Kulik et al., 2004), the syntactic parsers that can provide part-of-speech labels but also syntactic dependencies between words, such as Stanford dependencies parser (nlp.stanford.edu/software/stanford-dependencies.shtml) (Marneffe & Manning, 2008), semantic role labelers that annotate predicate-argument structure in the text with labels assisted through lexical knowledge bases such as NomBank (nlp.cs.nyu.edu/meyers/NomBank.html) (Meyers et al., 2004), PropBank (http://verbs.colorado.edu/~mpalmer/projects/ace.html) (Palmer, et al., 2005), FrameNet (framenet.icsi.berkeley.edu/fndrupal/) (Baker et al., 1998), all tend to substantiate the premise that the conversion of scientific language to its formulaic expression—such as the one described in Harris's paper—can be easily obtained from a corpus of scientific articles. However, although a number of advances have been made in this area, there is still no method for a full conversion of scientific texts into formulaic expressions and further into a sublanguage envisioned by Harris. Probably the largest obstacle that stands between a relatively simple, efficient, and domain-independent conversion of the language of scientific articles into their formulaic expressions concerns the problem of full understanding of the meaning of the document. This is certainly a hard barrier to overcome as the tools and resources that currently exist to aid this conversion can only recover or, better, intuit a portion of rather than the full meaning of a given text.

1.4 DISCOURSE DETECTION AND PROCESSING

The last two decades have seen tangible progress in the direction of converting the unstructured language of scientific articles into the structured format and the linguistic tools and resources previously mentioned certainly in aided these efforts. Tools and methods that support these kinds of inquiries are described in the following paragraphs. One example is the Claim Framework (2010), a domain independent annotation scheme which reflects how scientists communicate their findings. The Claim Framework focuses on five types of scientific claims such as explicit and implicit, correlations, observations and comparisons (Blake, 2010). The method developed by Blake (2010) carries the potential to allow information scientists to track the development of the discussion about one entity or one concept in literature and in this way analyze the shift in the vocabulary associated with the mention of this entity. Perhaps more important still, the methods allow the identification of different statement types in scientific literature and also the development of the statement and its subsequent reception through literature.

Another example is a study whose aim was to recognize discrete sentence categories from scientific articles. The identified categories, eleven in total, were the following: Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result, and Conclusion (Liakata et al., 2012). Recognizing these scientific artifacts in the corpus of articles in an automatic way can help us not only track the development of a hypothesis but also, and more crucially, compare objectives from different studies and their respective methods. The study by Liakata et al. (2012) has recently been amplified by Groza et al. (2013) to account for the recognition of core scientific artifacts in literature such as hypothesis, background, motivation, objective, and findings. The ultimate goal was to find supporting and contradicting statements for a hypothesis or summarizing findings that relate to a particular hypothesis. This automatic categorization of sentences helps us track the discussion about one entity or one biomedical concept and identify the paradigm shift (or the change in the perceptions) surrounding one concept or entity.

In a similar manner, the method proposed in this study allows us to track comparative statements made about one entity or one biomedical concept and identify how this entity compares to another entity and also the basis on which the entity was compared to another entity. The method proposed in this study relies on grammatical structure to indicate and summarize the information about one entity. However, given that it is possible not only to recognize comparative claims made about one entity but also explicit and implicit claims about one entity, the method described in this study will allow us to join comparative claims about one entity and in this way obtain a fuller picture about how an entity is discussed in a corpus of articles. The method proposed here also relies on the relations that exist between compared entities to identify a particular view of the text. Within the larger scheme of more general knowledge and more domain-specific knowledge basis, a comparative ontology does not strictly belong to either one (it can be both, depending on the discipline or genre). Comparative relations appear across domains and a broader study is needed to establish the frequency and the manner of expressing comparisons in different domains. In this study the focus is on biomedical texts and the relations extracted from these texts.

1.5 RELATIONSHIP EXTRACTION

Moldovan et al. (2000) make a distinction between more general and domain-specific knowledge bases such as Medline and the Genomes Online Database. Both more general knowledge databases and domain-specific databases are needed for knowledge intensive applications. At the present moment, however, general purpose knowledge databases (such as WordNet, for example) are both more common and more accessible than domain-specific ones. According to Moldovan et al. (2000), the development of domain-specific databases can be carried out in a manner similar to that witnessed in the field of general purpose databases. Specifically, the authors propose a method of establishing seed concepts, identifying noun phrases in which the concepts appear, recording the environment of the noun phrases while paying special attention to lexico-syntactic patterns that occur between nouns and, in time, using the patterns to extract patterns of the same kind in which at least one of the concepts appears as a constituent.

The gap that exists between the relationships encoded in general purpose and domain-specific knowledge bases is filled by domain knowledge. Domain-specific knowledge is specialized, and the type and accessibility of material differ significantly from what is found in more general applications. Freebase contains approximately a half million instances of a more general type of relationship such as book/author. The situation is quite different with more specialized relationships: they are neither as omnipresent (if we take into account journal subscription walls that provide barriers to information) nor as frequent. Data sparsity issues as well as the lack of domain-specific knowledge bases are the two main barriers to identifying and extracting relationships from a scientific corpus. For example, the medical domain is characterized by a host of specialized relationships for which no database with encoded relationships exists. To illustrate, Rosario & Hearst (2004) identify the following semantic relationships between Diseases and Treatment: Cure (Treatment cures Disease), Prevent (Treatment prevents the disease), Side Effect (Disease is a Result of a treatment), and No Cure (Treatment does not cure Disease). A comprehensive and extensive database (or databases) of fine-granular relationships such as the ones described above still needs to be developed. The Unified Medical Language System (UMLS) can be considered to be an exception to this rule but even within the UMLS—similarly to WordNet—the ISA relationship is identified as "the most primary link between the most semantic types" (http://www.nlm.nih.gov/research/umls/new_users/online_learning/SEM_004.html). The UMLS also encodes non-hierarchical relationships that indicate physical, spatial, temporal, functional and conceptual relationships between concepts and is well-known for Parent-Child (broader/narrower relationships). Rosario & Hearst (2004), however, used MeSH to map the words to their semantic categories rather than for training their model or verifying the results.

Regardless of these limitations found in specialized domains, the methods reminiscent of the ones developed by Hearst (1992), Girju (2006), Mintz et al. (2008) have been applied to the scientific domain. One example of such work is the study conducted by Rosario & Hearst (2004) which was described earlier. This study was subsequently elaborated by Frunza et al. (2011) – they used the same annotated set of diseases and treatments but different data representation. Rindfleisch et al. (2000) use medical knowledge bases such as MEDLINE to extract drug and gene names and to establish a relationship between genes and drugs used in cancer treatment. Their method relies on syntactic information between a gene and a drug but also on the rules that can help establish the connection between drug and genes.

Frunza et al. (2011) refer to three major approaches to extracting relationships in sciences: co-occurrence analysis, rule-based approaches, and statistical methods. The co-occurrence methods assume that a relationship exists between two entities if they occur in the same sentence and sometimes the relationships is inferred through the presence of some keyword. These methods are known to provide a good recall but poor precision (Frunza et al., 2011). Rule-based approaches usually rely on syntactic or semantic information extracted from the text and require significant human effort in devising the rules. The rule-based approaches can be divided into syntactic rule-based approaches and semantic rule-based approaches. Statistical approaches employ a variety of algorithms and data representation techniques such as bag-of-words, part-of-speech information, syntactic dependencies, as well as semantic labels to learn the environment in which the concepts appear.

What connects the methods for extracting relationships in the science domain to the methods developed by Hearst (1992), Girju (2006), Mintz et al. (2008) is that they all operate on the level of the sentence. Generally speaking, the methods described focus on the identification of two concepts in the same sentence and on recording the environment around these concepts. While this approach seems reasonable it typically does not solve the anaphoric reference issue – for example, an instance where a particular concept or entity are referred to in the sentence but are not named explicitly. The patterns identified in such a way may thus be ambiguous; however, semantic information about the concept may be able to help with the task of disambiguation. That said, semantic information or a specialized knowledge base may not be readily available in more specialized domains, which would then have a restrictive effect on the application of the methods described above. As mentioned previously, the UMLS ontology is one notable exception that assigns semantic categories to biomedical concepts; however, even this ontology fails to account for the more specialized, specific or more fine-granular relationships.

The methods described in this study also operate on the level of the sentence and do not venture to capture and synthesize information that spans sentences. Although some information is inevitably lost when we do not account for anaphoric references, the resolution of anaphoric references in comparison sentences is something that future work can consider.

1.5.1 Comparison relation

As indicated earlier, identification of comparison sentences has been a focus of a number of studies. Less work has been done in the area of identifying main predicates of comparison sentences such as compared entities, the nature of the relationship that binds them and also the basis of comparison, or the aspect/endpoint based on which the entities were compared. One exception, however, exists: the SemRep tool, an interactive tool that identifies the main predicates in the medical literature (Rindflesch & Fiszman, 2003). A sub-group of relations is dedicated to the so-called comparison type relations such as: *compared_with*, *lower_than*, *higher_than*, and *same_as*. Additionally, the tool recognizes the entities that participate in a comparison relation. The main difference between the approach described in this study and the SemRep is that the entities are not matched against MetaThesaurus before the predicate relation is identified. While, to a certain extent, this is a limitation of this method because such matching would allow deeper inferences to be made, at the same time, it can be argued, this is an advantage of the method described in this study: the entities and relations are not predefined and thus bounded by the limits of a knowledge base meaning that the approach presented here can be used on collections where ontologies are not available.

Another important difference between the SemRep tool and its functionalities and the method described in this dissertation lies in the output. The SemRep tool provides a semantic triple (subject, predicate, object) —the basic data structure of the Semantic Web (Berners-Lee et al., 2001). However, the basic data structure of the Semantic Web, a semantic triple, is not an adequate for representing the main predicates of a direct comparative sentence explored in this work. The main facets of a direct comparative sentence that the method elaborated here aims to identify and extract are compared entities (minimum two), the basis of comparison, and the result of the comparison or the relationship that binds the two compared entities and shows whether a property common to both entities has *increased, decreased*, *remained the same*, or was *different*. Thus, a semantic triple is not an adequate structure for representing the interaction between the compared entities, the endpoint and the main predicate. An N-quad (https://www.w3.org/TR/n-quads/), a quadruple (Damerow, 2014), “contextualized triple,” or a Named graph allow the addition of more context to the semantic triple. In the case of a direct comparison—the main focus of this study—the endpoints or the basis on which the compared entities have been compared can represent the contextual component for the relation that connects subject, object and predicate. The Quadriga System, developed by Damerow (2014), allows researchers to add the relations extracted from the text to a graph structure, and represent the data obtained (new knowledge) through a network. Although storing and representing the information from direct comparison sentences is not the focus of this study, n-quads look more promising for representing the information extracted from comparison sentences than a semantic or a database triple.

Another difference that sets apart the method as delineated in this dissertation from the SemRep tool is the comparison relation coverage. The method described here is bounded by a list of 70 comparison anchors that are used to recognize a comparison sentence in the text. While this list is by no means exhaustive and does not cover all the comparison relations, these 70 anchors account for more expressions of comparisons than the SemRep tool can currently recognize. More partcularly, the SemRep tool does not provide good coverage for non-gradable comparisons: for example, relations such as *similar to* and *different from* are currently not recognized. In this respect, the methods elaborated here have the potential to diversify the number and types of comparison relations currently distinguishable by the SemRep tool. For example, when the following sentence, recognized in this study as a comparison sentence, is run through SemRep, the *comparison relation* that is semantically inherent to the sentence is not recognized. The *location_of* and *part_of* predicates were recognized whereas *indistinguishable from* did not qualify as a comparison relation. Interestingly, the SemRep tool also recognized *compared_to* as a predicate of the semantic triple. In this work, *compared_to* represents an anchor that indicates that the sentence is most likely a comparison sentence but *compared_to* is always an *assumed* predicate of the method in this study and never the main relation. The assumption of the direct comparison sentence is that the two entities have been compared to each other. The fact that they have been compared thus is an assertion statement that would not necessarily be stored in a database or represented through the graph structure. What is of interest is the result of the comparison and the change that has occurred or not occurred as the result of comparing two or more entities. This represents another difference between the method as outlined in this study and the SemRep tool.

To conclude, this dissertation complements the list of comparative relations in the SemRep. Once a sentence is distinguished as a comparison sentence based on the presence of certain lexical and syntactic clues and based on the presence of a comparison anchors, the actual predicate and the nature of a comparison relation between two or more entities are identified. Relationship identification and the rules that are used to infer the type of relation that binds the compared entities are described in Chapter 5 of this dissertation. Four general types of gradable and two non-gradable types of comparison sentences have been identified. This typology is based solely on the syntactic and lexical features that determine how a comparative relation is expressed in a sentence.

1.6 TEXT SUMMARIZATION

Although this dissertation only provides what can be referred to as a prototype of a multi-document summary (See Section 2.11 “Proto-type of a multi-document summary”), the extracted facets from comparative sentences have the potential to be used as the background information for a text summarization system. Given that the gold standard for a multi-document summary that uses facets extracted from comparative sentences in the background as envisioned by this work does not exist and given that this dissertation calls for the creation of such standard, a few words should be mentioned about text summarization in general and about text summarization evaluation methods in particular. Given that the creation of such standard would not be an easy task the following discussion aims to elucidate the issues surrounding this task and also suggest viable alternatives to the creation of a gold standard.

Broadly defined, summarization evaluation method can be divided into extrinsic and intrinsic ones. The main difference involves the task of summarization. Extrinsic evaluation methods assess the usefulness of the summary for a given task, whereas intrinsic evaluation methods estimate its quality. Intrinsic evaluation methods can further be divided into content evaluation and text quality evaluation (Steinberger & Jezek, 2007). Content and quality text evaluation are distinct, though related, concepts. Content evaluation is focused on establishing how well the created summary covered the main topics or main content of the source text while quality text evaluation is focused on readability, grammar and cohesion of the created summary (Steinberger & Jezek, 2007). According to Halteren & Teufel:

> […] the best way to evaluate a summary is to try to perform the task for which the summary was meant in the first place and measure the quality of the summary on the basis of degree of success in executing the task (2003).

Although this represents the best way to evaluate the summary, Halteren & Teufel proceed to conclude that this is such a time-consuming task to set up that it “cannot be used for day-to-day evaluation needed during system development” (Halteren & Teuefel, 2003). Because of this inherent difficulty, intrinsic evaluation is a more common type of evaluation with automated text summarization tasks. Something that connects both extrinsic and intrinsic text summarization evaluation methods is the comparison of the results against the summary constructed by humans or against an “ideal” summary. However, the concept of an “ideal summary” is problematic. It shares certain characteristics with the judgment of relevance in the Information Retrieval field. All of the factors that determine and have an impact on relevance determination, such as cognition factors, situational, motivational, and user background factors, play a role in establishing and creating an “ideal” summary. And yet, despite these limitations, it has been established that “the relative effectiveness of different retrieval strategies is stable despite marked differences in the relevance judgments used to define perfect retrieval” (Voorhees, 2000). The concept of an “ideal summary”, however, still cannot claim the same level of reliability. Typically, the selection of sentences for inclusion in a summary suffers from a lack of both inter- and intra-subject reliability. Specifically, as discussed in a study by Resnick (1961), “the lack of inter and intra-subject reliability seems to imply that a single set of representative sentences does not exist for an article. It may be that there are many equally representative sets of sentences which exist for any given article.” Resnick’s conclusion refers to an inherent flaw or weakness of any text summarization evaluation that uses human-generated summaries or the content selected by humans either as the gold standard for building a summarization system or during the evaluation phase of the system itself. Salton et al. (1997) explored information contained in manual extracts and found that, on average, the information that is regarded as most important by one person would overlap with the information that is viewed the same by another person 46% of the time. This rather low overlap carries deep implications for any text summarization system that uses either a gold standard created by the humans or a manually constructed summary to judge against the automated system. Additionally, Lin and Hovy (2002) found that human judges agreed with each other no more than 82% of the time, based on a sample of 5,291 total judgments on the single document summarization evaluation task and about 92.4% in 6,963 total judgments on the multi-document summarization task. Similar to the results of the study by Rath, Resnik and Savage (1961), this last example indicates that not only do humans differ between each other in the ranking and evaluation of the quality of the created system (inter-rater reliability or reproducibility), but they do not agree with each other (intra-rater reliability or stability). This is the reason why Lin & Hovy emphasize that one should treat with caution any interpretation of performance figures that ignores this instability of manual evaluation (2002).

In general, gold standards, including gold standards created for text summarization tasks, take a long time to build and are costly to realize. It was reported that it took NIST creators 3,000 hours to create a gold standard for the 2001 and 2002 Document Understanding Conferences (Lin & Hovy, 2003). An additional disadvantage is that gold standards are usually made for a particular information task only, which then restricts the possibilities for their reuse.

Gold standards in the text summarization domain can be divided into those built primarily for evaluation purposes and those to be used for both training and test (evaluation) purposes. An example of the latter gold standard was described in a study conducted by Teufel & Moens (2002). The study, working with a corpus of 80 conference paper articles in the field of computational linguistics, aimed to automatically identify sentences in several defined categories: Aim, Textual, Own, Background, Contrast, Basis and Other. It was established that the identification of these rhetorical categories can help place the conference paper articles within a larger context and establish linkages that connect a given paper with earlier similar papers. One notable finding of the study was a comparatively low precision and recall on the categorization of sentences achieved by a human annotator. When compared against the results provided by Annotator B, Annotator C achieved only 50% precision and 54% recall for Contrast and 82% precision and 34% recall for Basis category. These two categories were subsequently the hardest to predict for humans, system, and baseline method (Teufel & Moens, 2002). Although related, recognizing sentences that belong to a certain type or category is not quite the same as the task of recognizing the most salient sentences or most informative sentences to be included in a summary. Although these results primarily indicate a low level of agreement about the Contrast and Basis category in the corpus of scientific articles, they also indicate a low level of understanding of what Contrast and Basis categories imply and what kind of sentences they tend to identify. This rather poor result of identifying certain categories of sentences in the text by human judges has an influence on how well the system can recognize the sentences. Or, put differently, it is difficult for a system that is trained to perform automated sentence classification to do well if the agreement about and understanding of categories in question is rather low among human evaluators.

Though not without its challenges, the gold standard that was created in the study conducted by Teufel & Moens (2002) provided a highly useful model not just for the training component but also for its automated evaluation. In general, any automated evaluation of a text summarization system is less costly, uses fewer resources, and can be performed more efficiently than evaluations conducted by human judges. However, automated evaluation of the content of the text summary comes with certain costs, compared with manual assessment. One example of an automated text summarization evaluation system is the ngram overlap method, developed by Lin & Hovy (2002). This method estimated the level of ngram overlap between the model summary and the system summary. The weakness of this model, however, was that it did not take into account synonymous phrases. The Pyramid automated evaluation method (Nenkova et al., 2007) was subsequently developed to take into account the problem of synonymous expressions. This method used a richly annotated gold standard that checked the source text for the appearance of phrases and content units that express the same content and assigned them a score based on the frequency a given phrase appeared in the text. Yet even with this innovation in place and available to researchers in the field, the summaries that concentrate on the information content captured generally pay less attention to the coherence and cohesion or grammar of the text. The reason for this is that these two tasks are hard to coordinate within a single setting. It should be noted, however, that if the summarization task is focused on reducing the original content and keeping the original sentences intact, then the grammar of individual sentences will not be compromised. And yet, this reduction does not guard against a diminished or hampered sentence flow and general coherence.

An example of a groundbreaking study that aimed to paraphrase a sentence in such a way that the sentence remains grammatical and at the same time retains the most important information is that of Knight & Marcu (2000). The results were evaluated with regard to the quality of the text (grammar) *and* the text’s informativeness (meaning, the most salient features of the sentence are retained). The only exception here was that the summarization unit was not a short or long text or even a paragraph but a single sentence. Additionally, the study used human judgments (rather than an automated method) to evaluate the results and assign scores from 1 to 5, depending on how well the decision-based compression algorithm managed to capture grammar and identify the most important words in the sentence. The overall results indicated that the decision-based model performed rather well, though semantics and the meaning of the sentence were sometimes negatively affected by the application of the compression algorithm. The final conclusion was that humans performed better than the algorithm but the performance of two compression algorithms was much closer to humans than to the baseline (Knight & Marcu, 2000). This study represents an important first step in the direction of evaluating both the content and the quality of the text or (other) multiple aspects of the text, albeit on the sentence level.

The examples provided above indicate that a full understanding of a document continues to elude automated text summarization systems. The systems typically rely on a summary or on multiple summaries that were created by humans and that serve as the gold standard in the automated evaluation of text summarization tasks—a concept susceptible to the vagaries (or more precisely, the inherent subjectivity) of human judgment. As we saw from the example of the Teufel & Moens study (2002), humans do not always have a full understanding or, equally important, a shared understanding of the categories they are annotating. These factors, in tandem, tend to limit the performance of the system.

The strengths of automated evaluation methods that focus on informativeness rather than quality, such as the ngram overlap method (Lin & Hovy, 2003), include their relatively easy implementation and also an efficient method of evaluation of content overlap. However, these methods typically do not take more complex aspects such as cohesion, coherence, anaphoric reference, sentence flow, and grammar, into account. Creating an automated evaluation of text summarization systems that can appropriately gauge and correctly weigh both the content and the quality of the text remains a challenging task.

The methods described in this study identify the compared entities, endpoint, and the predicate or the semantic relation that binds the three entities. While the chief premise of this dissertation is that the main elements of direct comparative sentences viewed in aggregate have the potential to summarize scholarly literature, this dissertation does not aim to create a text summarization tool. Given that the gold standard for identifying these entities and relations does not exist and given that creating such standard would require a considerable amount of resources, a somewhat easier and more manageable method of evaluating the output involves focusing on several entities and analyzing the information aggregated about these entities by reporting the precision, recall, and F_1 measure. In the chapters that follow, several such analyses are conducted where a sample of sentences is extracted and analyzed in terms of achieved precision, recall, F_1 and accuracy measures with respect to the identified roles. The results of this work, however—as indicated in the Conclusion chapter of this dissertation--highlight the need for the evaluation standard that can be used for improving and building on the method outlined herein. Such an annotated corpus might not only improve the precision and recall with which comparative facets are extracted from sentences but it might also positively influence precision and recall of any text summary that uses extracted information from comparative sentences in the background.

Chapter 2 will demonstrate how the comparative sentences are identified in scholarly articles and the precision and recall at which such sentences are extracted. This chapter also provides a prototype of a tabular multi-document summary by focusing on the drug metformin and identifying other drugs and interventions with which this drug has been compared, the basis of comparison or the endpoint, and also the result of the study.

Chapter 3 then delves into how the model in the pilot study was improved and demonstrates heuristics/rules and analysis that have been conducted to test model improvement. The distribution of different facets in the collection of biomedical scholarly articles is discussed and a typology of comparison sentences based on their semantic characteristics is provided.

Chapter 4 describes the addition of a new collection of scholarly articles—those dedicated to and focused on breast cancer—as a test collection to indicate how well the improved models perform when applied on a previously unseen collection.

Chapter 5 focuses on identifying the main predicate of a comparative relation. Four gradable and two non-gradable types of comparison relations have been identified based on the collection that is used in this thesis. In the discussion concluding the chapter, the results of how well the method can recognize individual claims and negation in comparative sentences are also included. Final chapters, Chapters 6-8, are dedicated to the discussion of implications and limitations of this work and also its conclusion.

CHAPTER 2: PILOT STUDY

2.1 IDENTIFICATION OF COMPARISON SENTENCES[3]

The first step of the task that aims to identify and extract main semantic elements of comparative sentences is the identification of *comparative sentences* in the text. Earlier work suggests that comparison phrases (such as *compared with*) provide good recall (98%), but the resulting precision can be low (32%) (Jindal & Liu, 2006). In the current setting, recall is much more important than precision because candidate sentences identified are subsequently processed using machine learning to identify the specific role that each noun plays in the comparison. Thus the operational system would be tuned for optimal recall during comparison sentence extraction and then for optimal precision in the second phase when the role of a noun phrase and the main predicate of a comparative relation is predicted. In the second phase of this process, the sentences that do not contain a noun phrase that is identified as either a compared entity or the endpoint may be eliminated.

To identify comparative sentences, a collection of adjectives and lexico-syntactic patterns was used. For example, a set of comparison marker phrases was developed that work well for both gradable (such as *compared with*) and non-gradable (such as *similar to*, *different from*) comparative sentences. Each transition phrase (either adjective or verb) has at least one corresponding preposition. For example, *different* can be followed by either the preposition *from* or *than.* Verbs such as *increase*, or *reduce* are good indicators of a gradable comparative sentence. The system for recognizing a comparative sentence used also a list of change terms (Blake, 2010), which includes the comparative agreement relation from the SPECIALIST Lexicon from the Unified Medical Language System (UMLS, umlsks.nlm.nih.gov, version 2014AB). Sentences containing lexico-syntactic paths that include either an UMLS adjectives, or the terms *better*, *more*, *less*, *worse, fewer*, and *lesser*, followed by the preposition *than* were also tagged as a candidate comparison sentence. Dependencies from the Stanford Parser (version 3.2)

[3] The text included in this chapter also appears in the following publication:

Blake, Catherine, & Lucic, Ana. (2015). Automatic endpoint detection to support the systematic review process. *Journal of Biomedical Informatics*. doi: http://dx.doi.org/10.1016/j.jbi.2015.05.004.

(Klein & Manning, 2003) allow the system to detect candidate sentences where the preposition does not immediately follow the adjective (or additional terms), as is illustrated in sentence 11 where text *diabetic subjects* that fills the entity 1 role occurs between the words *higher* and *than*.

(11) Fasting glucose [endpoint] was *higher* [relation] in diabetic subjects [entity 1] (168.8 55.2 \ mg/dl) *than* in nondiabetic subjects [entity 2] (93.9 9.6 mg/dl).$_{12882937}$

The previous lexico-syntactic paths work well for gradable comparisons. Additional paths that include adverbial modifier (advmod) or a finite clause subordinate (mark) were also used with the preposition *than*. This work extends the 35 features described in (Park & Blake, 2013) where comparative sentences were identified using three different classification algorithms: Naïve Bayes, Support Vector Machines and a Bayesian network. A machine learning approach was considered during the first step, but eventually it was established that the accuracy was sufficient without an additional layer of training.

2.2 COLLECTION STATISTICS

Using the method and the rules described in section 2.1, candidate comparative sentences were identified from more than 2 million sentences from full-text of the articles published in the journals *Diabetes, Carcinogenesis, and Endocrinology* and included in the TREC 2006 Genomics collection of scholarly articles (http://trec.nist.gov/pubs/trec14/papers/GEO.OVERVIEW.pdf). The journal *Diabetes* had the greatest proportion of comparative sentences (23,088 sentences, 5.41% of the total collection). *Carcinogenesis* had a slightly smaller number of candidate comparative sentences but almost the same proportion of comparative sentences (20,151, 5.01% of the total collection), and *Endocrinology* had the largest overall number of sentences because more articles were available (63,995 sentences, 5.35% of the total collection). As Table 1 indicates, 5.31% of sentences in the collection were identified as candidate direct comparatives, which is consistent with 5.11% reported in prior work (Blake, 2010).

Comparative sentences from all three journals were similar with respect to the number of noun phrases, which was 8 on average. Given that the prediction is made on the noun level, the possibility of correctly assigning the role by chance is 1/8=0.125 or 12.5%. Even if the sentence had fewer candidate nouns, 4 for example, the chances of correctly predicting the role by chance would not exceed 0.25 or 25%. Thus, 25% precision can be regarded as the *baseline* precision for this task. The number of words (as opposed to noun phrases shown in Table 1) in the direct comparisons ranged between 5 and 129. Sentences that contained fewer than three noun phrases were considered for removal, however, since several sentences used the same noun phrase to describe both entity 1 and endpoint only sentences that contained 1 noun phrase were removed (Table 1 indicates collection statistics after sentences that contained only one noun phrase had been removed).

Journal	Articles	Sentences	Candidate sentences	Noun phrases Avg (min, max)	% candidate sentences
Diabetes	2,142	426,694	23,088	8 (2, 35)	5.41%
Carcinogenesis	1,958	394,835	20,051	8 (2, 58)	5.01%
Endocrinology	5,100	1,194,114	63,995	8 (2, 46)	5.35%
Total	9,200	2,015,643	107,134		5.31%
Sentences with 1 or 2 comparison anchors and <= 2 change anchors and <= 40 words			86,864		4.31%

Table 1: Summary statistics from three full-text journals from the TREC Genomics collection

The number of comparison anchors varied between 1 and 8, and the number of change anchors was between 0 or 1 and 12. A closer inspection of sentences that contained many comparison anchors and change anchors revealed that many were not direct comparisons, thus only sentences that contained 1 or 2 comparison anchors and 0, 1 or 2 change verbs were considered in subsequent steps, which was the majority of the candidate sentences (102,260/107,134, 95.5%).

With respect to sentence length, candidate comparative sentences drawn from *Diabetes* and *Carcinogenesis* contained an average of 30 words and sentences from *Endocrinology* had an average of 29 words, which is longer than sentences in news stories which have on average 23.4

words (Bojar et al., 2014). Sentences were further constrained to include 40 or fewer words, which again was the majority of sentences (89,355/107,134, 83%). Applying both the number of change and comparison anchors and sentence length constraints resulted in 86,864 (81%) candidate comparative sentences that were considered in the consequent step (see Figure 1).

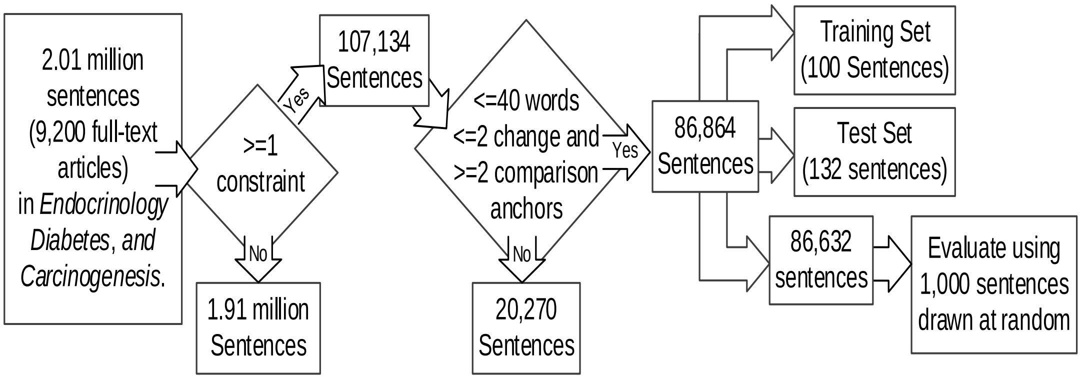
2.01 million sentences (9,200 full-text articles) in *Endocrinology Diabetes, and Carcinogenesis.*
>=1 constraint
Yes
No
107,134 Sentences
1.91 million Sentences
<=40 words <=2 change and >=2 comparison anchors
Yes
No
20,270 Sentences
86,864 Sentences
Training Set (100 Sentences)
Test Set (132 sentences)
86,632 sentences
Evaluate using 1,000 sentences drawn at random

Figure 1: Experimental design for Pilot study

The accuracy of comparison sentence identification was established by randomly selecting 1,000 from the 86,632 sentences that were not part of either the training or test set. Out of 1,000 sentences, 866 sentences were direct comparative sentences, thus the precision of this step was 86.6%. Establishing recall is difficult when such a small proportion of sentences (5.31%) contain a direct comparative, so recall is explored using a situated evaluation described in section 2.9.

2.3 DISTRIBUTION OF COMPARATIVE FACETS IN SENTENCES

To better understand the structure of comparative sentences and to be able to infer which candidate noun phrase in the sentence fulfills the role of compared entity 1, compared entity 2 and the endpoint, the distribution and position of the three comparison facets were analyzed in 100 candidate comparison sentences. Candidate sentences were extracted from the TREC Genomics collection. Based on whether the first entity in the comparison sentence appeared before the term that indicated a change (the change term) and based on whether the endpoint appeared before the transition words that indicate comparisons and contrasts (comparison anchor) four types of comparisons were identified.

The first type comprised the sentences in which the first entity was positioned before the verb that indicated a change as well as before the endpoint that appeared after the change verb. The following is an example of the sentence that follows this pattern:

(12) Short-term treatment [entity 1] of ovariectomized rats [entity 1] with estradiol plus progesterone [entity 1] caused *significantly* [relation modifier] *decreased* [relation] preoptic Gal-R1 mRNA levels [endpoint] compared with those after treatment [entity 2] with estrogen only [entity 2]. 9751492

In this sentence, the first compared entity that consists of three nouns phrases, *ovariectomized rats, and estradiol plus progesterone*—occurs before the verb that indicates change (change verb)—*decrease*—and before the basis of comparison (basis of change)—*preoptic Gal-R1 mRNA levels*. The structure that was characterized by the entity 1 that was followed by the change verb that was followed by the endpoint that occurred near or close to comparison marker term was the most frequent pattern in the set of 100 sentences. In this sentence, the first compared entity (entity 1) truly has a role of an agent, an active inducer of change in entity 2. However, given that this type of sentence represents only one type of comparison sentence and given that the first entity does not always have the role of a change inducer, a more neutral definition of compared entities as entity 1 and entity 2 was selected. Entity 1 corresponds to the entity that is mentioned first in the sentence and that typically occurs before the comparison anchors such as, compared to and similar to. Entity 2 most often occurs after the comparison anchor: hence the label entity 2.

The second type of comparisons was characterized by a reverse ordering of the first compared entity and the endpoint: endpoint preceded the first compared entity and the first compared entity preceded the comparison marker phrase. Here is an example of a sentence that followed this pattern:

(13) Buserelin-stimulated serum testosterone levels [endpoint_A] in male G_11_alpha knockout mice [entity 1_A] was *significantly* [relation modifier] *higher* [relation_A] than in control mice [entity 2_A], although Buserelin stimulated estradiol levels [endpoint_B] in female G_11_alpha knockout mice [entity 1_B] were *lower* [relation_B] than in control mice [entity 2_B]. 9607776

The third type involved the change verb that preceded the endpoint and first entity. Here is an example of a sentence that followed this pattern:

(14) Furthermore, a *trend toward a lower* [relation] hepatic microsomal free cholesterol [endpoint_A] and triglyceride concentrations [endpoint_B] was observed with atorvastatin [entity 1] compared with simvastatin treatment [entity 2]. 28200735

The change term in this sentence (*lower*) precedes the endpoints (*hepatic microsomal free cholesterol and tryglyceride concentrations*) and the first entity (*atorvastatin*).

The final pattern involved the sentences in which the comparison marker phrase occurred towards the beginning rather than towards the end of the sentence. For example, in the following sentence:

(15) Thus, *similar* [relation] to norepinephrine [entity 2_A] and epinephrine [entity 2_B], dopamine [entity 1] in the presence of IL-1beta [entity 1] *induced* [relation] a synergistic stimulation [endpoint] of IL-6 release [endpoint]. 9927320

the comparison marker phrase (*similar to*) occurs at the beginning of the sentence rather than towards the end of the sentence. As in earlier patterns, however, the comparison marker phrase is followed by the second compared entity *(norepinephrine*, *epinephrine*).These four surface patterns render themselves to the following representation:

Type I comparison: ENTITY 1 – CHANGE VERB – ENDPOINT – COMPARISON MARKER – ENTITY 2
Type II comparison: ENDPOINT – ENTITY 1 – CHANGE VERB – COMPARISON MARKER– ENTITY 2
Type III comparison: CHANGE VERB – ENDPOINT– ENTITY 1 – COMPARISON MARKER – ENTITY 2
Type IV comparison: COMPARISON MARKER – ENTITY 2 – ENTITY 1 – ENDPOINT

Figure 2: Characterizing surface level constructs of comparison sentences

One pattern that starts to emerge from this representation is the close association of the second compared entity with comparison marker that typically occurs towards the end of the sentence or comparative claim. One exception, however, is the fourth pattern in which the comparison marker and the second compared entity occur towards the beginning rather than toward the end of the claim although even with this pattern they remain closely associated.

Another pattern that this representation revealed was the close association of the first compared entity, the verb that indicates a change that has occurred (in gradable comparisons), and the endpoint. These three elements of comparison sentences are frequently closely associated although their distribution and position in relation to each other varies.

In contrast to type I and III, the fourth sentence structure (IV) does not include a change verb because the comparison is non-gradable (as opposed to gradable). However, although the change verb is absent first entity still maintains close association with the endpoint even in non-gradable sentences. As we will see later, this close association is one of the reasons why it is difficult for the classifier to separate the role of the entity from the role of the endpoint.

The results of this analysis informed a number of features that were used in a supervised model that aimed to predict the role of two compared entities in the sentence as well as the endpoint based on which they were compared.

2.4 MACHINE LEARNING METHOD

To determine in an automated way the noun phrases that satisfy the roles of compared entities and the endpoint from candidate comparative sentences, machine learning was used. Candidate noun phrases are generated automatically using the domain independent dependency parser (Klein & Manning, 2003) (version 3.2). The system was provided with both single words (*raloxifene*, *ERalpha* and *estradiol* in sentence 16) and complex compound noun phrases (*RBA assay* and *highest affinity*) (Blake & Rindflesch).

(16) In this RBA assay, raloxifene [entity 1] exhibited the ***highest*** [relation modifier] ***affinity*** [relation] for ERalpha [endpoint] relative to estradiol [entity 2]. 10579349

Two classifiers, the support vector machine (SVM) and Generalized Linear Model (GLM) were built for each role. The Oracle Data Miner (ODM) 11g, Release 2 implementation of the algorithms were used, where the linear kernel and complexity factor of 0.167 was used with the SVM and a confidence level of 0.95 was used with the GLM based on our experience with the ODM in other experiments (http://www.oracle.com/webfolder/technetwork/tutorials/obe/db/11g/r2/prod/bidw/datamining/ODM11gR2.htm).

2.5 FEATURE DESCRIPTION

Twenty-six features were used for the recognition of the first compared entity whereas for the second compared entity and the endpoint a subset of twenty-one features was used. Features were informed by prior work (see related work) and by a training set comprising 656 noun phrases (in 100 sentences) that were drawn from three different full-text journals (*Diabetes, Carcinogenesis and Endocrinology*). Most features can be grouped into one of the following anchor categories.

A. Change anchors are typically associated with gradable comparisons, such as *minimize*, *lose* and *accelerate*. The system uses a lexicon of 770 verbs that were modified from (Blake, 2010).

B. Comparison anchors such as *similar to, different from*, and *compared with* identify both gradable and non-gradable comparisons. At this stage, the system used 65 comparison marker phrases that were modified from (Park & Blake, 2013). Within this set, the system distinguishes between two types of comparison anchors, those in which:

   i. The first word in the phrase is immediately followed by the preposition (i.e there is no gap between the first word in the phrase and the preposition)

and those in which:

ii. The first word in the phrase is immediately followed by the noun (i.e. there is a gap between the first word and the preposition)

C. Evidence anchors are verbs that indicate a finding. The system used a set of 432 evidence-based verbs such as *acknowledge, result, imply, view, find, illustrate* that were created for this experiment.

The change (A) and comparison (Bi,Bii) anchors have been explored in the comparative mining. In contrast, the evidence anchors (C) have been explored with respect to scientific rhetoric (Teufel & Moens, 2002) but evidence terms have not been explored with reference to identifying entities in the contexts of how the entities are compared.

The first set of features measure the raw and syntactic distance between each candidate noun and the change, comparison and evidence anchors. Raw distance is the number of words that occur between a noun phrase and the anchor. Sentences containing more than one anchor of the same type are assigned the distance to the closest anchor and values of 1000 are assigned to sentences without an anchor. Type Bi and Bii are captured separately for raw distance, thus there are eight features that capture the raw distance before and after each anchor, which are used in all three classifiers (entity 1, entity 2 and endpoint).

Unfortunately, raw distance is sensitive to the number of words in a noun phrase and conjunctive clauses, which are frequently employed in technical writing. Syntactic distance employs the dependency representation of a sentence to mitigate against the raw distance limitations, where syntactic distance is the number of nodes between the head noun of a noun phrase and each anchor that appears in the same branch of the dependency tree. As with raw distance, the minimum distance is used if multiple anchors of the same type are used in a sentence.

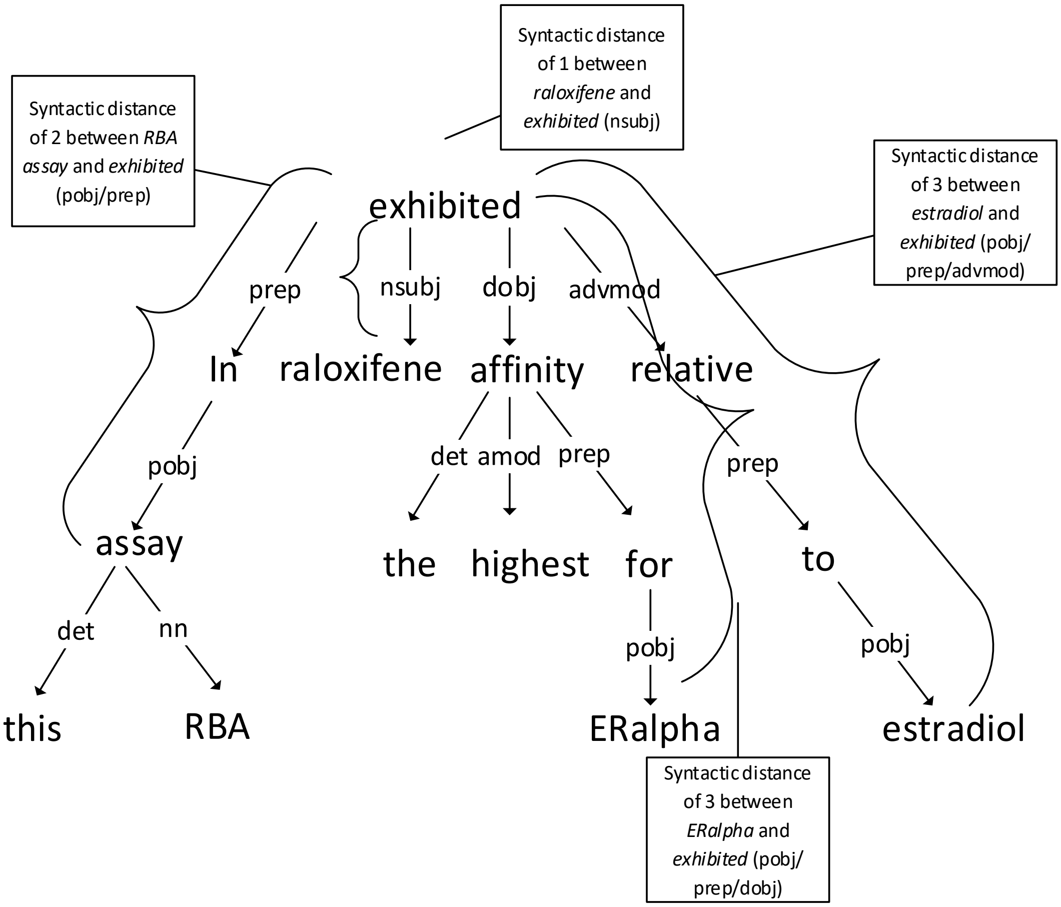
Syntactic distance of 1 between *raloxifene* and *exhibited* (nsubj)
Syntactic distance of 2 between *RBA assay* and *exhibited* (pobj/prep)
Syntactic distance of 3 between *estradiol* and *exhibited* (pobj/ prep/advmod)
exhibited
prep
nsubj
dobj
advmod
In
raloxifene
affinity
relative
pobj
det
amod
prep
prep
assay
the
highest
for
to
det
nn
pobj
pobj
this
RBA
ERalpha
estradiol
Syntactic distance of 3 between *ERalpha* and *exhibited* (pobj/ prep/dobj)

Figure 3: Raw and syntactic distances (sentence 16)

Figure 3 shows the syntactic dependencies generated from sentence 16 that contains the evidence anchor *exhibited*. The syntactic distance between each of the following nouns *RBA assay*, *raloxifene*, *affinity*, *Eralpha*, and *estradiol* and the root of the sentence exhibited is 2, 1, 1, 1, and 3 respectively. In addition to the evidence anchor, the noun phrase that fulfils the object role in this sentence (*estradiol*) has a syntactic distance of 1 with respect to the comparison anchor *relative to*. None of the other noun phrases occur in the comparison anchor branch of the dependency tree and would thus be assigned a distance of 1000. The second set of features capture the syntactic distance between the each noun phrase and each of the anchors (one for type A, Bi, Bii and C). These features are used in each classifier (SVM and GLM) for all three roles (entity 1, entity 2, endpoint).

The third set of features also employs the dependency structure, but rather than use the numerical distance, categorical features are used to capture the dependency path between a noun phrase and each anchor. For example, the path between the basis *RBA Assay* and the evidence anchor is pobj/prep. Similarly the path between the first compared entity *raloxifene* and the evidence anchor is nsubj. To avoid overfitting—making the model too complex, i.e., too fitted to the idiosyncrasies of the set of sentences that is used for training—two additional categorical features were included, one that captures the first 2 syntactic dependencies from the anchor and another that captures the first 3 syntactic dependencies from the anchor. These two features were used as more general features that would show less variation than all the dependencies connect the candidate noun phrases to each of the three anchors. A total of six categorical features capture the syntactic path between the root of the syntactic tree and each candidate noun phrases and each of the anchors (A, B or C).

The fourth set of features capture words that occur immediately before and after each noun phrase in the sentence, and the fifth feature(isLeaf), is set to true when the noun or noun phrase appears as a terminal node in the syntactic tree (a leaf). For example, in sentence 10 the head noun *controls* is not a leaf node but isLeaf is set to true because the modifier *lean* is a leaf. In contrast, isLeaf would be false for the noun *serum*.

The 21 features described thus far were used in each of the SVM and GLM classifiers to classify nouns that filled the endpoint or entity 2 roles. Initial experiments with entity 1 role suggested that an additional five features would improve performance (for a total of 26 features). Analysis of the 100 sentences in the training set revealed that entity 1 noun phrases were frequently a nominal subject in a sentence. The nominal subject feature is set to true when the dependency path contains a nominal subject (*nsubj*) and the noun is close to the evidence term (the syntactic distance after an evidence term is less than or equal to two). The boolean complement feature is set to true when the noun is part of a clausal complement (where nouns are more likely to be the first compared entity). Nouns that occur near a comparison marker anchor are more likely to be the second compared entity (entity 2). A feature (isFarFromComp) is set to *true* when the raw distance between the noun and the comparison anchor was greater than 20 in order to hone in on nouns that play entity 1 role. Lastly, two features that capture when the noun occurs between the comparison marker and the subsequent preposition are included in entity 1 model.

(17) A *twofold* [relation modifier] *increase* [relation] in ALT activity [endpoint] was observed in serum from diabetic rates [entity 1] compared with the lean controls [entity 2].15277384

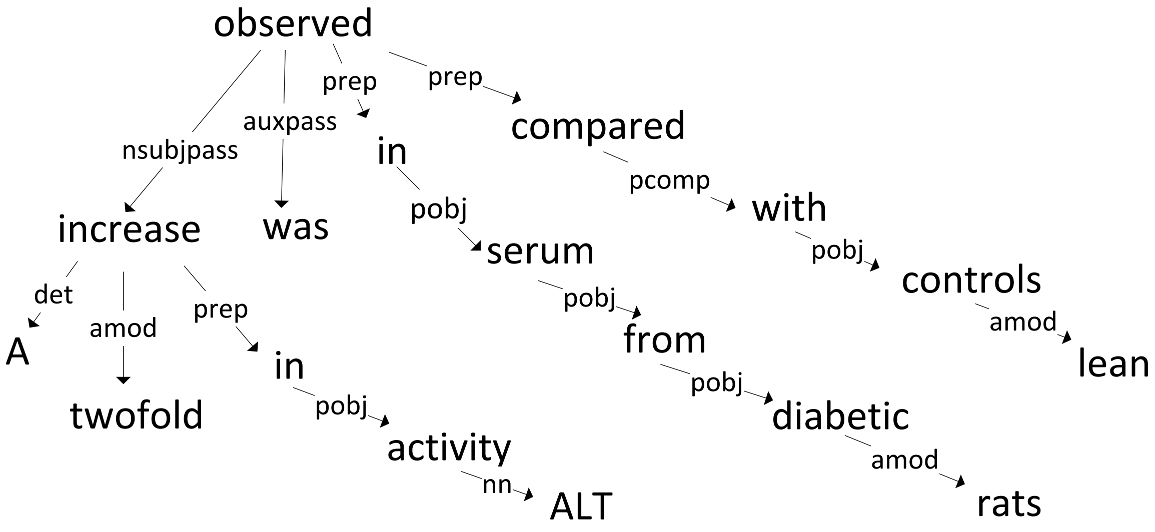

Figure 4: Syntactic dependency tree (sentence 17)

2.6 RESULTS

A training set of 656 noun phrases (in 100 sentences) was annotated by the author of this dissertation and her dissertation advisor. Differences between annotators were discussed until agreement was reached on the role a noun phrase played in the sentence as well as on the boundary of the noun phrase. The features were established using prior work and annotations in the training set. A test set of 936 noun phrases in 132 sentences was created by selecting 132 sentences at random from the 86,764 candidate sentences identified in step 1 that were not part of the training set (see Figure 5).

The classification models created for the training set were then applied to the test sentences. Three different support vector machine (SVM) and generalized linear models (GLM) were built, one for each role: entity 1, entity 2 and endpoint. The classification performance of the SVM was better than the GLM on the training set with respect to all measures, precision, recall, F_1 measure, and accuracy for all three facets (see Table 2).

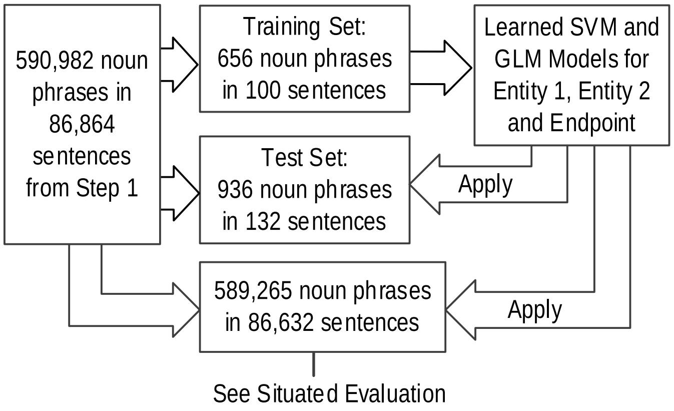
590,982 noun phrases in 86,864 sentences from Step 1
Training Set: 656 noun phrases in 100 sentences
Learned SVM and GLM Models for Entity 1, Entity 2 and Endpoint
Test Set: 936 noun phrases in 132 sentences
Apply
589,265 noun phrases in 86,632 sentences
Apply
See Situated Evaluation

Figure 5: Experimental design

As with the training set sentences, the SVM model outperformed the GLM model in the test set, but the overall results for entity 1 and endpoint were lower in the test than the training set (see Table 2). Despite the additional features, all metrics except for GLM accuracy had lower performance when predicting entity 1 role. Performance for entity 2 was similar to the training set, but the 0.05 difference between the GLM and SVM models in the training set was reduced to 0.01 in the test set. The GLM and SVM models showed similar performance for all roles.

	Entity 1		**Endpoint**		**Entity 2**	
	GLM	**SVM**	**GLM**	**SVM**	**GLM**	**SVM**
Precision	0.50 (0.46, 0.54)	0.59 (0.55, 0.63)	0.57 (0.53, 0.61)	0.67 (0.63, 0.71)	0.66 (0.62, 0.70)	0.78 (0.75, 0.81)
Recall	0.75 (0.72, 0.78)	0.96 (0.94, 0.98)	0.84 (0.81, 0.87)	0.94 (0.92, 0.96)	0.91 (0.89, 0.93)	0.99 (0.98, 1.00)
F_1 Measure	0.60 (0.56, 0.64)	0.73 (0.70, 0.76)	0.68 (0.64, 0.72)	0.78 (0.75, 0.81)	0.76 (0.73, 0.79)	0.87 (0.84, 0.90)
Accuracy	0.79 (0.76, 0.82)	0.85 (0.82, 0.88)	0.82 (0.79, 0.85)	0.88 (0.85, 0.91)	0.9 (0.88, 0.92)	0.95 (0.93, 0.97)

Table 2: Test set results showing 95% confidence intervals

Technical writing, such as the prose in scientific articles is more difficult to read than non-technical writing such as news stories, in part because the sentences are longer and thus tend to have more clauses and more noun phrases. In order to see the impact of sentence length on the system performance, the test sentences were partitioned into shorter (containing 30 or fewer words) and longer sentences that contain more than 30 and less than or equal to 40 words (see Tables 3 and 4). Sentence length had little or no impact on accuracy, where performance dropped by 0.05, 0.02, and 0.01 for entity 1, endpoint and entity 2 respectively. However there was a drop in F_1 performance (0.16) for both entity 1 and endpoint models and a small drop in F_1 performance for entity 2 prediction (0.01). Longer sentences appeared to have had a greater negative effect on the GLM models than the SVM models.

	Entity 1		**Endpoint**		**Entity 2**	
	GLM	**SVM**	**GLM**	**SVM**	**GLM**	**SVM**
Precision	0.38 (0.35, 0.41)	0.38 (0.35, 0.41)	0.40 (0.37, 0.43)	0.42 (0.39, 0.45)	0.78 (0.75, 0.81)	0.74 (0.71, 0.77)
Recall	0.63 (0.60, 0.66)	0.58 (0.55, 0.61)	0.69 (0.66, 0.72)	0.64 (0.61, 0.67)	0.75 (0.72, 0.78)	0.8 (0.79, 0.81)
F_1 Measure	0.47 (0.44, 0.50)	0.46 (0.43, 0.49)	0.51 (0.48, 0.54)	0.51 (0.48, 0.54)	0.76 (0.73, 0.79)	0.77 (0.74, 0.80)
Accuracy	0.72 (0.69, 0.75)	0.73 (0.70, 0.76)	0.71 (0.68, 0.74)	0.73 (0.70, 0.76)	0.92 (0.90, 0.94)	0.91 (0.89, 0.93)

Table 3: Test set results on shorter sentences (<=30 words) showing 95% confidence intervals

	Entity 1		Endpoint		Entity 2	
	GLM	SVM	GLM	SVM	GLM	SVM
Precision	0.30 (0.26, 0.34)	0.33 (0.29, 0.37)	0.32 (0.28, 0.36)	0.35 (0.31, 0.39)	0.76 (0.72, 0.80)	0.71 (0.67, 0.75)
Recall	0.56 (0.52, 0.60)	0.55 (0.51, 0.59)	0.62 (0.58, 0.66)	0.59 (0.55, 0.63)	0.72 (0.68, 0.76)	0.77 (0.73, 0.81)
F_1 Measure	0.39 (0.35, 0.43)	0.41 (0.37, 0.45)	0.43 (0.39, 0.47)	0.44 (0.40. 0.48)	0.74 (0.70, 0.78)	0.74 (0.70, 0.78)
Accuracy	0.70 (0.66, 0.74)	0.72 (0.68, 0.76)	0.69 (0.65, 0.73)	0.73 (0.69, 0.77)	0.92 (0.89, 0.93)	0.91 (0.89, 0.93)

Table 4: Test set results on longer sentences (> 30 and <= 40 words) showing 95% confidence intervals

In summary, both the GLM and SVM classifiers were better able to predict entity 2 than the endpoint or entity 1 across all measures (precision, recall, F_1 measure and the overall accuracy).

2.7 INTERPOLATED PRECISION AND RECALL

The results presented in previous section employed thresholds determined from the training set, but thresholds can be adjusted to favor precision or recall. Figure 6 shows the precision-recall curves for entity 2, endpoint and entity 1 for the both the GLM and SVM models. Features were created based on earlier work and the training set, so the results shown in figure 6 should be interpreted as an upper bound of system performance. The overall test set is also shown, along with the results from the short and long sentences in order to show how sentence length impacts system performance. The precision reported is the highest precision at each of the 11 recall levels i.e. P interpI=max r'≥p(r') as defined in (Manning et al., 2008, p. 145).

Figure 6 indicates that nouns fulfilling entity 2 role are easier to predict. The GLM and SVM models show similar performance, with a maximum recall of 0.8 and a maximum precision of 0.8 for short sentences. At the highest recall level (0.8), precision was 0.7 for longer sentences using the SVM model. The application of GLM model resulted in the highest recall level of 0.7

and the maximum recorded precision of 0.8. Compared to the interpolated precision-recall graph for entity 2 identification, the endpoint models show a drop in precision for long sentences. For short sentences, the maximum precision at the highest recall level (0.8) was 0.5 and with the all sentences set, at the highest recall level (0.7), the maximum precision recorded was 0.5. With the SVM model, the highest recall level using the short sentences set was 0.7 and the maximum precision achieved at this level was 0.5.

Entity 1 role was the most difficult comparison facet to predict. Similar to the endpoint, the GLM model achieved a slightly better performance on the short sentences set compared to SVM model. The highest recall was 0.7 for short sentences, where the maximum precision 0.5. Compared to SVM performance using the same set the highest recall level was 0.6 and the maximum precision 0.5. The overall set of test sentences had a maximum precision at the highest recall level (0.6) of 0.4 in both the GLM and SVM models. Entity 1 role was the most difficult to predict in long sentences. The application of SVM model on this set resulted in the highest recall level of 0.5 and the maximum precision of 0.4. The GLM model applied on the overall test set resulted in the highest recall level of 0.6 and the maximum precision of 0.3.

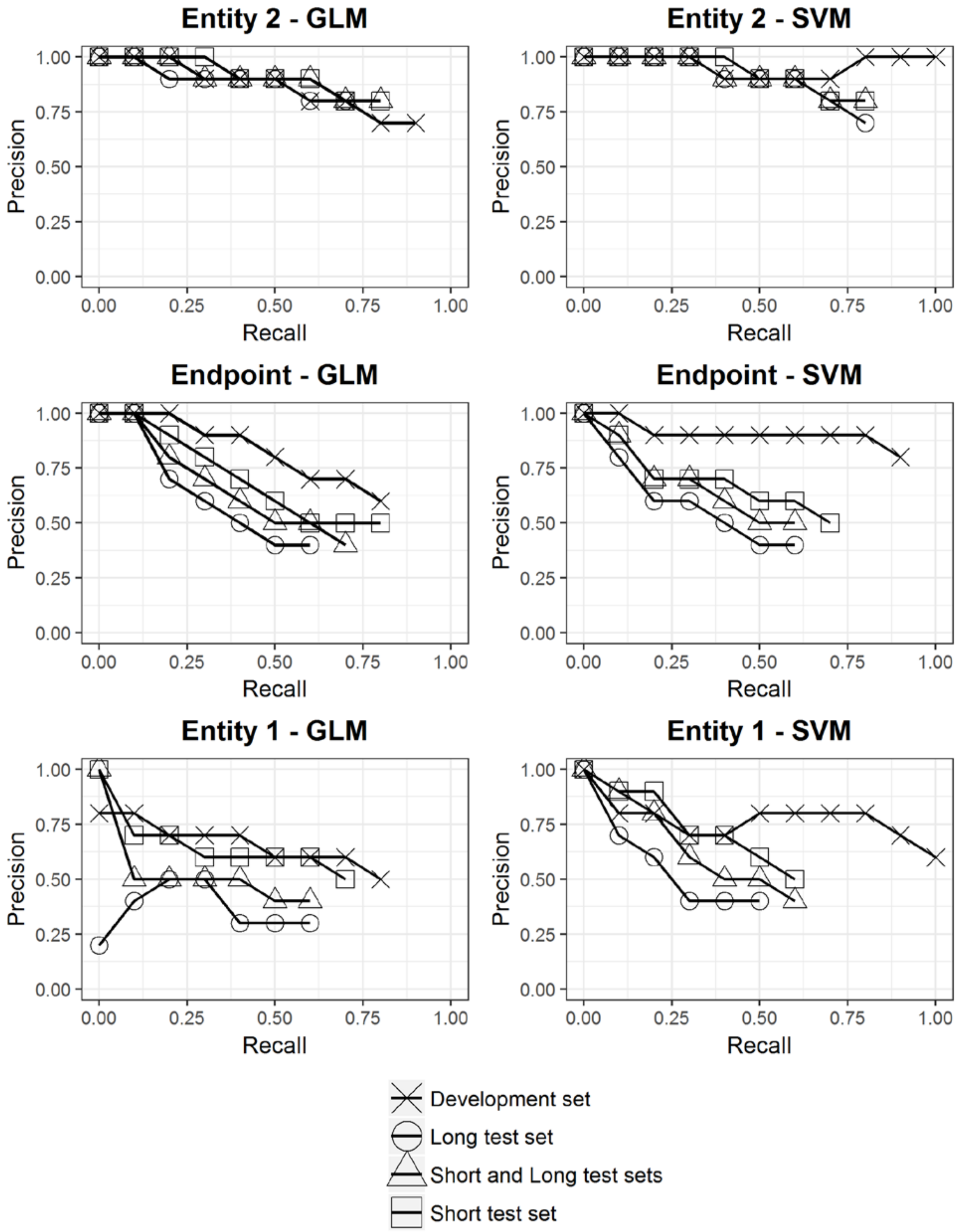

Figure 6: Interpolated precision and recall for entity 2, endpoint, and entity 1

2.8 ERROR AND FEATURE ANALYSIS

A random sample of 90 noun phrases (30 entity 1, 30 entity 2, and 30 endpoints) from each predictive model for each journal (10 in each journal for each of the three roles) were selected for further analysis. The average recall for all three models was 0.82 and the average precision was 0.58. A closer inspection of features that played a major role in the model showed that both entity 1 and endpoint are in close proximity to the change anchor, particularly for gradable comparisons. However, because *both* entity 1 and endpoint are near the change anchor, the predictive model incorrectly labels first compared entity as endpoint and vice versa. The confusion matrix in Table 5 shows that entity 1/endpoint misclassification caused all but 1 of the recall errors (an endpoint that was misclassified as entity 2).

		PREDICTED				
		Entity 1	**Endpoint**	**Entity 2**	**Total**	**Recall**
ACTUAL	**Entity 1**	15	4	0	19	0.79
	Endpoint	4	13	1	18	0.72
	Entity 2	0	0	24	24	1.00
	Other	10	13	5	23	
	Total	30	30	30	90	

Table 5: Error analysis of the SVM model for compared entities and the endpoint

With respect to precision, model results would need to be further processed to identify the correct noun phrase in the 5 cases where the model predicted an anaphoric reference, and in 3 cases where the model predicted a noun phrase that introduced the first compared entity. Complex sentence structures with multiple comparisons lead to 3 additional errors and preprocessing errors at either a sentence or noun phrase level caused an additional 9 errors when the role of the candidate noun phrases was identified. Lastly, 6 errors during the machine learning method step when the noun role was predicted were propagated from errors in the step when a sentence was incorrectly tagged as a comparison.

Table 6 summarizes the root cause of precision errors.

Factors that influence precision	Entity 1	Endpoint	Entity 2
Endpoint identified as Entity 1 or Entity 2	4		1
Entity 2 identified as Entity 1	1		
Entity 1 identified as Entity 2		4	
Anaphoric reference	3	1	1
Introduces the role	3		
Complex comparison structure	2		1
Preprocessing errors – sentence level	1	2	
Preprocessing errors – noun phrase level		4	2
Not a full comparison		2	
Not a comparison	2	4	1
Total errors	15	17	6
Precision	0.50	0.43	0.80

Table 6: Factors responsible for the loss of precision in the SVM model

Several features contributed to the model. The six categorical features that captured the syntactic path between the root of the syntactic tree and the noun phrases were particularly informative, where syntactic paths that were most indicative of the first compared entity (from the root of the tree) were passive nominal subjects, either a conjunction or preposition following a nominal subject, a direct object or a conjunction following a direct object and a direct object following a causal complement were the strongest features. Syntactic paths that were indicative of the endpoint (again with respect to the root of the dependency tree) were a nominal subject, an open causal complement followed by a preposition, and a passive nominal subject followed by a preposition.

The second compared entity was the easiest to identify, where noun phrases that occurred between 1 and 5 words after the comparison anchor were frequently the second compared entity, for example a sentence containing “compared to the control group” would have a distance of 2. Object noun phrases were also likely to be in a path that started with a prepositional complement following the comparison anchors (pcomp/pobj/dep/acomp/prep/pobj/conj, pcomp/pobj/prep/pobj/prep/pobj/advmod, or pcomp/pobj/prep/pobj/partmod/dobj) or that started with a prepositional object (prep/pobj/prep/pobj/conj). The presence of conjunctions found in these informative paths highlights the complexity of sentences found in the genre of scientific articles. The most informative syntactic path for the second compared entity starting from the root of the sentence was prep/conj.

The existing manual systematic review processes requires that each team member extract facts from each article independently and then reconcile difference. The system that is envisioned in this dissertation would be tightly integrated into that manual processes, where the system would make predictions that would then be verified before moving into the analysis stage. The goal with the situated evaluation is to demonstrate how facets from direct comparison sentences provide insight into experimental basis. Once the system evaluation reported earlier was complete, the noun (and noun phrases) that played any role (compared entities or the endpoint) were ordered with respect to the number of times that each phrase appeared. Metformin, an antidiabetic drug used to treat diabetes, appeared frequently and thus became the focus for this situated evaluation. The purpose of this evaluation is not to measure the most frequent noun phrase in the sample of articles that were in TREC because in a real systematic review much more care would be required to ensure that articles from all the most relevant journals were selected, but rather to explore the impact of redundancy, to demonstrate that comparison facets can be detected automatically without prior knowledge and to show how the facets extracted from this system can be used to identify areas where the literature agrees, disagrees and where there are gaps.

2.9 SITUATED EVALUATION

To evaluate identification of comparative sentences, all candidate comparative sentences (a set 86,864 sentences) containing either Metformin or the brand names Glucophage, Glumetza, Fortamet, or Glucophage XR were identified. Metformin appeared in 1,178 sentences from all three journals in the following forms: *metformin*, *active metformin*, *antecedent metformin treatment*, *intensive lifestyle metformin*, *metformin group, metformin mouse*, *metformin treatment*, *metformin-treated group*, *metformin-treated hepatocyte*, *metformin-treated mouse*, *metformin-treated rat, metformin-treated subject*.

This situated example illustrates the gap between the way in which a drug intervention is discussed in a scientific article and the way in which a drug would be represented in an ontology. Although you would expect metformin to be in a drug ontology such as those in the Unified Medical Language System (UMLS), it is unlikely that the phrase *metformin-treated group* would be captured. Similarly *metformin-treated mice* reflects a drug, an experimental process, and a group and it would be odd to have the noun phrase *metformin-treated* in a drug ontology. One of the other frequent nouns that emerged from the test journals was *diabetic*, which is in the UMLS, but *nondiabetic* is not in the UMLS and nor should the latter term be added. These results suggest that there is a gap between the terminology used to capture concepts in an ontology and the terminology used in scientific discourse. It is not just a question of domain coverage, but rather an inherent issue of surface level differences.

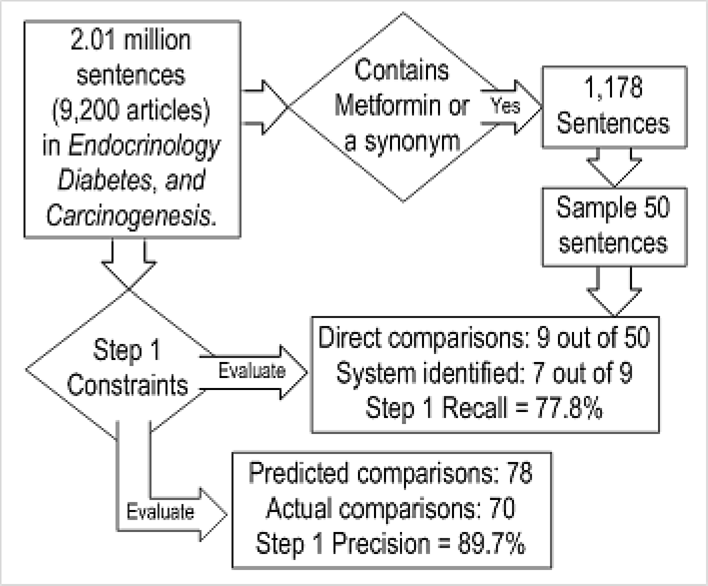
2.01 million sentences (9,200 articles) in *Endocrinology Diabetes, and Carcinogenesis.*
Contains Metformin or a synonym
Yes
1,178 Sentences
Sample 50 sentences
Step 1 Constraints
Evaluate
Direct comparisons: 9 out of 50
System identified: 7 out of 9
Step 1 Recall = 77.8%
Evaluate
Predicted comparisons: 78
Actual comparisons: 70
Step 1 Precision = 89.7%

Figure 7: Situated evaluation design for the identification of comparison sentences

A random sample of 50 sentences was drawn from the 1,178 candidate sentences that contained at least one mention of Metformin or a synonym (see Figure 7). Of the 9 direct comparison sentences found, the system correctly identified 7 providing a recall of 77.8%. The two missed sentences (18 and 19) are shown below.

(18) Fasting serum insulin concentrations [endpoint] *decreased* [relation] *significantly* [relation modifier_A] and *similarly* [relation modifier_B] during both rosiglitazone [entity 1] and metformin therapy [entity 2] by 4±1 and 4±2 mU/l, respectively (Fig 1). 15277403

(19) Water intake [endpoint] was randomly monitored throughout the study, was found to *increase* [relation_A] in proportion to body weight (r = 0.69, P < 0.001), and was *not* [relation modifier] *different* [relation_B] among treatment groups (0.093 ± 0.004, 0.098 ± 0.011, and 0.098 ± 0.004 ml · g^{-1} · day^{-1} for control [entity 1_A], metformin-[entity 1_B], and rosiglitazone-treated mice [entity 1_C], respectively). 15983227

The proximity of the change and comparison anchor was responsible for missing sentence 18, and the system appears to have limitations when the author presents the results in complex parenthetical structures. Please not that in sentence 19, the comparison anchor is *different among*, which does not lend itself well to entity 1 and entity 2 differentiation. The remaining sentences contained metformin, but metformin did not play the role of a compared entity or the endpoint in a comparison.

Further analysis was conducted to better understand how redundancy in scientific articles would impact overall system performance.

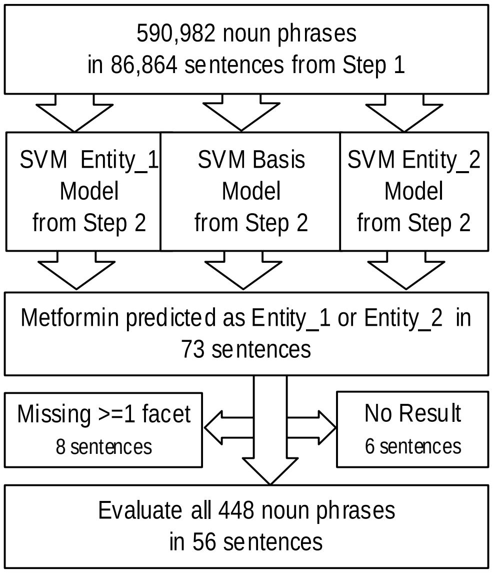
590,982 noun phrases
in 86,864 sentences from Step 1
SVM Entity_1 Model from Step 2
SVM Basis Model from Step 2
SVM Entity_2 Model from Step 2
Metformin predicted as Entity_1 or Entity_2 in 73 sentences
Missing >=1 facet
8 sentences
No Result
6 sentences
Evaluate all 448 noun phrases
in 56 sentences

Figure 8: Situated Evaluation

All 1,178 candidate comparison sentences from step 1 were searched and a total of 16 total sentences were found in which metformin played the role of a compared entity or the endpoint but the sentences were missed by the classifier. To evaluate noun role identification the sentences in which the system predicted that Metformin acted as the compared entity or the endpoint were closely inspected (see Figure 8). Of the 73 sentences, 6 (8%) contained noun phrases that filled the role of a compared entity or the endpoint but did not actually report a result. No system constraints are currently in place to remove non-result sentences, so this evaluation provides information about how much precision might be improved if such constraints were in place. Eight additional sentences (13.7%) were missing one of the comparison facets, which were necessary to complete the tabular summaries in tables 10-12. Metformin did actually play a role in 56 of the sentences, providing a precision in of 76.7%.

All 448 nouns phrases in the 56 sentences that the system identified and where Metformin did actually play a role were manually annotated and compared with the system predictions. The results (shown in Table 7) are consistent with experiments reported in earlier sections, in that the comparison entity 2 was identified with the highest precision (0.88) recall (0.88) and accuracy (0.95). The SVM model achieved better performance than the test set across all metrics and all facets (entity 1, endpoint and entity 2). This suggests that direct comparison sentences that include at least one drug may be more easily structured than those that report other types of results. The

GLM model also performed better than the test set for entity 1 and endpoint (with the exception of recall for the endpoint, which was 0.03 lower than the test set), however the GLM model in the situated evaluation was not as good with respect to predicting noun phrase that played an entity 2 role. The SVM model outperformed the GLM model across all metrics.

In contrast with the earlier test results that differed by only 0.01, the differences between the GLM and SVM models were more pronounced in the situated evaluation, but the differences were still small, where accuracy differed by 0.02, 0.05 and 0.06 for entity 1, endpoint, and entity 2 respectively. The SVM model provided much higher F_1 measures than the earlier test set for comparison facets and the GLM model was higher for entity 1 and endpoint, but about the same for entity 2. These differences in results may be due to the smaller sample size of 448 noun phrases (in 56 sentences) rather than 936 noun phrases (in 132 sentences) in the test set, or the system may be leveraging regularities in how authors compare drugs.

	Entity 1		Endpoint		Entity 2	
	GLM	SVM	GLM	SVM	GLM	SVM
Precision	0.45 (0.40, 0.50)	0.50 (0.45, 0.55)	0.53 (0.48, 0.58)	0.62 (0.58, 0.66)	0.75 (0.71, 0.79)	0.90 (0.87, 0.93)
Recall	0.65 (0.62, 0.68)	0.80 (0.77, 0.83)	0.59 (0.56, 0.62)	0.67 (0.64, 0.70)	0.75 (0.72, 0.78)	0.88 (0.86, 0.90)
F_1 Measure	0.53 (0.50, 0.56)	0.62 (0.58, 0.65)	0.56 (0.53, 0.59)	0.64 (0.61, 0.67)	0.75 (0.72, 0.78)	0.89 (0.87, 0.91)
Accuracy	0.77 (0.74, 0.80)	0.79 (0.76, 0.82)	0.78 (0.75, 0.81)	0.83 (0.81, 0.85)	0.89 (0.87, 0.91)	0.95 (0.94, 0.96)

Table 7: Situated evaluation in diabetes treatments showing 95% confidence intervals

The system had incorrectly missed 16 comparative sentences so the question then was how much information was in the sentences that were missed. The analysis showed that all of the information from 7 of the 16 sentences was in at least one of the other 56 sentences that were retrieved by the system. The 9 missing sentences contained three variations on existing aspects (*fat cell content, insulin release,* and *beta-cell function*), two new aspects (*adiponectin, FFA concentrations*), and two new compared entities (*anti hyperglycemic drug PE* and *TZDs).*

2.10 ERROR AND FEATURE ANALYSIS OF THE SITUATED EVALUATION

An error analysis was conducted for the Metformin situated evaluation. Table 8 summarizes the false positives and negatives for each model. The number of false negatives in the entity 1 model suggests that the threshold established with the training set could be lowered to improve overall performance. In contrast, the endpoint model had a similar number of false positives and false negatives, which suggests that the threshold was well set. Entity 2 also had a good threshold.

Entity 1		Predicted		Total		Endpoint		Predicted		Total		Entity 2		Predicted		Total
		Yes	No					Yes	No					Yes	No	
Actual	Yes	74	18	92		Actual	Yes	68	34	102		Actual	Yes	91	12	103
	No	74	282	356			No	42	304	346			No	10	335	345
Total		148	300	448		Total		110	338	448		Total		101	347	448

Table 8: Contingency tables for situated evaluation of entity 1, endpoint and entity 2

As with the error analysis described earlier, misclassified compared entities contributed to misclassification errors in the situated evaluation, where noun phrases that played either a compared entity or the endpoint role appeared in the same sentence location. The general order of the roles are entity 1-change anchor-endpoint, endpoint-entity 1-change anchor, change anchor-endpoint-entity 1 or entity 2-entity 1-endpoint. The same sentence structure can have either an entity 1 or an endpoint in the same place, so the current set of features do not have enough discriminatory power to distinguish between these facets.

In several cases the model predicted an introductory clause of a sentence rather than the first compared entity, for example the entity 1 model incorrectly predicted “conclusion” instead of “intensive lifestyle” for sentence 20 shown below. Removing such noun phrases before running the model would be one strategy to resolve this error, or adding additional post-processing that finds the following clause may also help to alleviate this issue.

(20) In conclusion, intensive lifestyle intervention [entity 1] *reduced* [relation] levels [endpoint] of nontraditional cardiovascular risk factors [endpoint] both relative to placebo [entity 2_A] and to a lesser degree relative to metformin [entity 2_B]. 15855347

Complex comparison structures also caused errors, particularly when the main focus of the sentence is not a comparison relation. Consider sentence 21 where the main focus is a comparison relation and then sentence 22 that compares *hepatic* and *renal responses* before and after treatment with metformin. Sentence 22 contains an additional claim that the antecedent metformin treatment did not have any influence on the results. This additional information is supplementary to the expression of the main comparison and as such introduces noise into the process of disambiguation of comparison facets that participate in the expression of a comparison relation.

(21) Interestingly, metformin [entity 1] (Fig 2C) also *inhibited* [relation] PTP opening [endpoint] with an efficacy *similar* [relation] to that of CsA [entity 2]. 15983220

(22) Nevertheless, since hepatic [endpoint_A] and renal responses [endpoint_B] observed in our subjects [entity 1] treated with metformin [entity 1] *did not* [relation modifier] *differ* [relation] from those not treated with it [entity 2], an influence of antecedent metformin treatment on our results seems unlikely. 12765948

One of the underlying premise of this work is that the fundamental unit of a comparison is a noun phrase, which is not always the case. Consider sentence 23, where the first compared entity really should be the "mechanism by which metformin activates AMPK". The current representation is inadequate to represent this compared entity. Similarly, the second compared entity in sentence 23 is also a mechanism that includes multiple noun phrases for the same entity, and would be approximated by the system by selecting more than one noun phrase. These relationships are captured in the Claim Framework as entity 1, entity 2 and endpoint modifiers, but modifiers are not implemented in the current system.

(23) Although the mechanism [entity 1] by which metformin [entity 1] activates AMPK [endpoint] remains unclear, it must be *different* [relation] from that of AICA riboside [entity 2], which acts by being converted to the AMP mimetic agent, ZMP. 12145153

Some system errors were caused by complex sentence structures, such as those that included the use of subordinating conjunctions, correlative conjunctions, and discourse terms such as *nevertheless*, *moreover*, *however* that authors use to describe nuanced experimental results. The model does identify syntactic structures with conjunctive clauses, however further pre-processing may help to alleviate some of the errors. Sentence 24 shows an example of how conjunctions can be layered within the sentence.

(24) Among all participants, including those who developed diabetes, fasting glucose [endpoint_A], insulin [endpoint_B], and proinsulin concentrations [endpoint_C] were significantly [relation modifier_A] *lower* [relation_A] than placebo [entity 2] at the first annual visit in the metformin [entity 1_A] and the lifestyle groups [entity 1_B] and *increased* [relation_B] during the 2nd and 3rd years, with the levels remaining *significantly* [relation modifier_C] *lower* [relation_C] than in the placebo group [entity 2] (Fig 2). 16046308

Sentence 24 contains 3 endpoints (*fasting glucose*, *insulin* and *proinsulin concentrations)* that all must be identified by the system to obtain perfect recall. In addition, the sentence contains more than one entity 2 (metformin and lifestyle groups) and again the entity 2 classifier would need to identify both noun phrases to achieve perfect recall. The real challenge comes with extracting the comparison claim that refers to the 2nd and 3rd year of intervention and the implicit reference to metformin in it. Lastly, this sentence highlights the importance of modifiers that are part of the claim framework, but not extracted in this dissertation.

2.11 A PROTOTYPE OF A MULTI-DOCUMENT SUMMARY

Systematic reviews typically synthesize evidence based on the same study design, such as randomized clinical trials in medicine or different animals in toxicology. The evaluation of 20 Metformin studies and the attempt to synthesize the evidence with respect to the number and type of treatments and interventions Metformin drug has been compared to took into consideration the medical subject headings (MeSH) of the articles. The MeSH headings were collected for each of the 20 different articles that contained the 56 verified Metformin sentences. All articles contained a MeSH of either Humans (16 articles) or Animals (5 articles). One articles had both Humans and Animals, but the majority of the findings and experiments reported in that paper focused on animals so the study was added to the Animals tabular summary.

Tables 9, 10, and 11 show tabular summaries of endpoints detected in human articles involving Metformin. Each table cell in the summary provides the result with respect to a given endpoint (shown as rows) that were used to compare Metformin with a given intervention (shown as columns). Columns 2-6 show drug comparisons and columns 7-9 show non-drug interventions. Endpoints identified by the system are listed in the leftmost column. The headings are abbreviated to just the first letter in the table cells, for example the T in row 1, column 2 refers to Troglitazone. The rightmost column shows how Metformin compares to a placebo or control group and shows that many of the endpoints measured with Metformin are frequently compared with placebo or control groups which is consistent with concerns raised by the comparative effectiveness community (Bojar, 2014).

The situated evaluation shows that in addition to measuring diabetes directly, insulin and glucose levels are measured to evaluate Metformin with other interventions (see Table 10). The tabular summaries shows endpoints in the original form, except for abbreviations where the full form of the abbreviation from the article is shown in the table to aid in readability. It may be possible for a system to automatically unify the endpoints, but we envision that the domain expert who is responsible for the systematic review would be directly involved in the decisions made concerning which rates, concentrations, sensitivities and suppressions should be combined. Because comparison sentences provide a densely packed summary of results users may need to return to the original article to determine which endpoints should be unified. For example the diabetes incident rate shown in the first row should only be unified with the risk for type 2 diabetes in the second row if the first article also refers to type 2 diabetes. Although the comparison sentence does not contain this information the system maintains a link back to the original study so that a user can accurately unify terms. In addition to the PubMed identifier that is shown in each cell of the tabular summaries in tables 10-12, the system maintains the specific sentence so that the user can go directly to the section of the article where the claim was made (space limitations prevented us from providing this additional level of detail in the tables).

The tabular summary that represents a prototype of a multi-document comparative summary provides insight into additional experiments that may be required, for example there is a gap in this collection with respect to measuring diabetes directly for several of the pharmacological treatments. These aspects were identified in the small set of 9,200 articles but the automated methods presented in this paper scale naturally to larger collections of full text articles to provide a more complete picture of endpoints that had been studied so that scientists and policy makers can obtain a better picture of where results differ and where there are gaps.

Diabetes						
Endpoint	Pharmacological treatment				Lifestyle	Placebo
	Rosiglitazone	Troglitazone	Metformin & Exanatide	Metformin, Sulfonylurea & Exanatide	Diet and exercise	
diabetes incidence rate		T < M* 15793255				
risk for type 2 diabetes					D more effective* 14633845	M < placebo 16046308
development of diabetes		T greater impact than M 15793255				

Table 9: Tabular summary of diabetes endpoint variations from human studies (* means significant)

Insulin						
	Pharmacological treatment				Lifestyle	Placebo
Endpoint	Rosiglitazone	Troglitazone	Metformin & Exanatide	Metformin, Sulfonylurea & Exanatide	Diet and exercise	
proinsulin concentrations						M <* placebo 16046308
Insulin						M < placebo 16046308
insulin action		T *> M (improvement) 11756319				
Insulin (concentrations)		T similar to M (decreased*) 11812753				M <* placebo 16046308
serum insulin concentrations	R <* M 15277403					
insulin sensitivity		T >* M (improvement) 15793255				
% supression by insulin	M < R 15277403					

Table 9: (cont.)

Glucose						
Endpoint	Pharmacological treatment				Lifestyle	Placebo
	Rosiglitazone	Troglitazone	Metformin & Exanatide	Metformin, Sulfonylurea & Exanatide	Diet and exercise	
Glucose		T similar to M (decreased*) 11812753				
fasting glucose						M <* placebo 16046308
hepatic glucose production during hyperinsulinemia	M < R 15277403					
glucose disposal		T greater efficacy than M 11756319				
glucose disposal (rate)		T > M, P<0.05 11812753 T > M 12606507				
serum fructosamine			M & E reduces > M alone 15331525	M, S & E reduces > M alone 15331525		
glycated haemoglobin (HbA1c)		T similar to M (decreased*) 11812753	M & E reduces > M alone 15331525	M, S & E reduces > M alone 15331525		

Table 9: (cont.)

Endpoint	Pharmacological treatment			Lifestyle intervention		Placebo
	Troglitazone	Rotenone	Cyclosporin A (CsA)	Intensive lifestyle intervention	Moderate-intensity treadmill running	Placebo / control
Phosphatidylinositol-4,5-bisphosphate 3-kinase (PI3K) activity	T > M 11812753					
AKT activity	T similar to M 11812753 T different to M 11812753					
AMPK-activated protein kinase (AMPK) alpha2					similar activation 12086935	
C-reactive protein (CRP)	T > (reduction) M 15855347					M < placebo 15855347 M <* placebo 15855347
complex I in human microvascular endothelial cells (HMEC-1)		M mild inhibitor compared with R 15983220				

Table 10: Tabular summary of remaining endpoints used to measure diabetes interventions in human studies (* means significant)

Endpoint	Pharmacological treatment			Lifestyle intervention		Placebo
	Troglitazone	Rotenone	Cyclosporin A (CsA)	Intensive lifestyle intervention	Moderate-intensity treadmill running	Placebo / control
permeability transition pore (PTP) opening			similar to M (inhibition) 15983220			
fibrinogen levels				L < M 15855347		
fat cells	T larger* M 11756319					
gene expression						no * differences 15855325
nontraditional cardiovascular risk factors				L < M 15855347		
hepatic and renal responses						no difference 12765948
nocturnal or postprandial lipolysis	M has no effect (in contrast to T) 12606508					
abdominal area						no changes by M 12540598
medication adherence	T > M 15793255					

Table 10: (cont.)

In contrast to Table 9 that shows endpoints with different representations, Table 10 shows endpoints that occurred only once in the human studies. Much of the intellectual work in a systematic review involves reconciling information that may seem contradictory. For example, pmid 11812753 reports that Metformin is both similar to and different from Troglitazone with respect to the endpoint Akt activity. In this case the user could return to the original sentences (shown below as 25 and 26) and would see that these findings are consistent with the experimental results. Sentences are drawn from anywhere in the article, and the sentence location suggests that sentence 25 refers to the conditions before the experiment and sentence 26 provides the actual result.

(25) The small effect [endpoint] of insulin [endpoint] to stimulate Akt activity [endpoint] before metformin treatment [entity 1] was *similar* [relation] to that in the troglitazone group [entity 2] before treatment (NS). 11812753

(26) However, in contrast to troglitazone treatment [entity 2], there was *no* [relation modifier] *enhancement* [relation] of Akt activation [endpoint] in response to insulin after metformin treatment [entity 1] (Fig 2B). 11812753

Endpoint	S 422	Benfluorex	Rosiglitazone	Pioglitazone	control	5-aminoimidazole-4-carboxamide riboside	AMPKK1 / AMPKK2
expression of glycolytic and gluconeogenic enzymes	S similar M 12145146	B similar M 12145146					
hepatic gluconeogenesis (mechanisms for reduction)		B markedly different to M 12145146					
AMP-activated protein kinase (AMPK)						A different to M 12145153	
metabolic effects						A very similar to M 12145153	
complex I inhibition			R inhibited (M didn't) 15047621	P inhibited (M didn't) 15047621			
respiratory control			R < M ↓15047621	P < M ↓ 15047621			
ADP-to-oxygen ratios with succinate			R < M ↓ 15047621	P < M ↓ 15047621			

Table 11: Tabular summary of endpoints used to measure interventions in animal studies (* means significant)

Endpoint	S 422	Benfluorex	Rosiglitazone	Pioglitazone	control	5-aminoimida-zole-4-carbo-xamide riboside	AMPKK1 / AMPKK2
upstream kinase							M might act differently AMPKK1 /AMPKK 12145153
peak fractional cell shortening (PS)					C unaffected by M 11334425		
islet amyloid prevalence (and severity)			R reduced* M 15983227				
proportion of beta-cell mass to islet mass					M >* C 15983227		
beta-cell mass					M < C 15983227		
mean islet mass					C not different to M 15983227		

Table 11: (cont.)

Endpoint	S 422	Benfluorex	Rosiglitazone	Pioglitazone	control	5-aminoimidazole-4-carboxamide riboside	AMPKK1 / AMPKK2
islet mass					C not different to M 15983227 M < C 15983227		
human islet amyloid polypeptide (hIAPP) contents					C not different to M (P=0.2) 15983227		
mouse islet amyloid polypeptide contents					C not different to M (P=0.07) 15983227		
pancreatic insulin content					M < C (P<0.05) 15983227		
Fasting plasma hIAPP					M < C (P<0.01) 15983227		
body weight					M < C 15983227		
fat mass					M <* C 15983227		

Table 11: (cont.)

Endpoints used in the five articles containing verified comparisons and involving studies with animals (i.e. the article had been assigned a MeSH of Animal) are shown in Table 11. The endpoints used in these experiments (see the first column) tend to focus on molecular and cellular activities that provide insight into the underlying mechanisms rather than the higher level endpoints measured in human studies. One of the advantages of using the system described in this paper is that a user does not need to fully articulate the endpoints shown in tables 10-12 in advance, but rather the endpoints are identified automatically by the system. To further emphasize this point, the following subsection will demonstrate the kind of results that can be expected if the nature and class of either of the entities that are compared or endpoint is specified in advance.

2.11.1 Noun mapping to an ontology

Given that this project uses a biomedical collection of articles and given that biomedical concepts are prevalent in this collection, mapping the text against the Unified Medical Language System (UMLS) knowledge base that includes a Semantic Network (McCray et al., 2001) of the concepts seems like a logical step. Several tools and resources allow such mapping of the free text to UMLS. MetaMap tool, for example, was developed by Lister Hill Center for Biomedical Communications which is part of Semantic Knowledge Representation (SKR). MetaMap is a program that is capable of analyzing biomedical text and mapping it against the UMLS Metathesaurus and in this way identifying biomedical concepts in the text. One of the important characteristics of MetaMap program is its variant generation characteristic through which identified nouns are expanded. This characteristic of the program is essential for mapping different representations of the same concept to Metatheasurus. MetaMap algorithm produces a score for each identified concept where a higher score indicates a higher confidence of the algorithm that the noun belongs to a certain concept. It is well-known that MetaMap recall is tightly connected to the UMLS coverage (Pratt, 2003) and so the concepts that are not in the UMLS will not be identified in the text. It is also well-known that MetaMap’s precision can be affected by variant generation process where the noun depending on how it is split can be matched against more than one concept. Fiszman et al. (2007) study described earlier was using the MetaMap tool.

Mgrep tool, on the other hand, was developed at the University of Michigan (Dai et al., 2008). A study that compared MetaMap and Mgrep (Shah, 2009) established that Mgrep outperformed MetaMap in all cases except in the case of recognizing Biological Processes in records from ClinicalTrials.gov collection. Mgrep is known as a fast and scalable tool and it has been used for building the Open Biomedical Annotator Web service (Shah, 2009).

The Open Annotator Web Service that was built by the National Center for Biomedical Ontology (NCBO) is yet another resource that allows the annotation of free text with the concepts that are in the UMLS but also with the concepts in biomedical ontologies outside of the UMLS. Back in 2009, the Open Annotator Web Service included concepts from 206 ontologies from the UMLS and the National Center for Biomedical Ontology (NCBO), 4,021,662 unique concepts, and 7,637,125 terms (Shah, 2009). At present, the Open Annotator Web Service includes more than 270 ontologies and terminologies (including all the vocabularies and terminologies that are part of the UMLS). The NCBO is a scientific organization charged with the task of bringing semantic technology to biomedicine (Musen et al, 2012). The most widely used tool created by NCBO is the Annotator tool that can annotate the natural language text and keywords with ontological terms from one or multiple ontologies (Musen et al., 2012).

A particular advantage that the NCBO Annotator tool has over MetaMap and Mgrep is that it allows annotation of the text with ontologies that are outside of the UMLS. Given that MeSH terms are particularly convenient for biomedical texts I used the Open Web Annotator Service to map 56 sentences that were identified by the system to contain Metformin drug as one of the compared entities to MeSH terms without restricting on the semantic category. The results are included in Table 12.

Metformin case study matching to MeSH terms					
	Full match	**Partial match**	**Not in the ontology**	**False match**	**Number of unique records**
Entity 1	36.36%	18.18%	44.44%	1.01%	99
Endpoint	32.95%	21.59%	43.18%	2.27%	88
Entity 2	52.73%	7.27%	38.18%	1.82%	55

Table 12: Percentage of unique terms mapped to MeSH terms using the Open Annotator Web Service (BioPortal)

As Table 12 indicates, four types of matches have been identified:

1) Full match
2) Partial match
3) Not a MeSH term
4) False match

A few words should be said about each of these categories.

1) Full match was identified as the match of the term or compound noun to a MeSH term that leaves little room for ambiguity between the term and the concept identified. For example, matching of the term insulin to insulin concept (id: http://purl.bioontology.org/ontology/MESH/D007328) leaves little or no room for ambiguity to occur. If the compound noun consisted of more than one word and each of the words was matched to a correct although distinct concept this was also considered a full match. For example, the compound noun *metformin treatment* was matched to the following concepts:

http://purl.bioontology.org/ontology/MESH/D008687 – Metformin
http://purl.bioontology.org/ontology/MESH/Q000628 – Therapy
http://purl.bioontology.org/ontology/MESH/D013812 – therapeutics

Although matching *metformin treatment*—the noun that consists of two words—to three concepts does not seem like an ideal solution, the combination of concepts still manages to imply the meaning of the compound *metformin treatment* correctly. More particularly, the concept *treatment* in MeSH is identified as a synonym of concept *therapy* which is an additional reason why this mapping was treated as as a full rather than partial match.

2) Partial match, on the other hand, was identified as the match that partially identifies the compound noun. For example, *fasting insulin concentrations* was matched to the concepts *Fasting*, *Insulin* and *Attention*. Although matches to *Fasting* and *Insulin* were correct the match to *Attention* is not applicable because the word *concentrations* does not refer to *Attention* but to a measurement value. This is the reason why this particular match was labeled as a partial match.

3) Terms that have not been matched to MeSH terms (*Not in the ontology* label) can be divided into two groups: those that seem too general or broad in nature or even too specific to be included in an ontology and those that seem like potential candidates to be included in the ontology. Among others, potential candidates included abbreviations that have not been matched to MeSH terms which suggests a need to first expand the abbreviations in the text and then do the match. Additionally, the terms such as *type 2 diabetes*, *placebo values*, *baseline values, control* or *controls, main antidiabetic compound* have also not been matched to MeSH vocabulary although they seem potential candidates.

4) There have only been a few false positives (False match), terms that have been wrongly associated with MeSH terms. One example is the compound noun *high concentration* that has been mapped to the concept *Attention*. *High concentration* in the context of the sentence refers to the concentration of the drug rather than to the general concept *Attention*.

As Table 12 indicates, approximately a third of the unique compound and single nouns that have been identified as entity 1 and endpoint (36.36% for entity 1 and 32.95% for endpoint) have been matched to MeSH terms using a full match method. Approximately 52% of unique terms that have been identified as entity 2 were fully matched to MeSH terms. Partial match accounted for approximately 20% of entity 1 and endpoints and only 7% of the entity 2 matches. Approximately 44% of entity 1 and endpoint of comparison have not been matched to any MeSH terms. This unmatched category can be divided into terms that truly do not belong in an ontology such as the terms that are too general in nature to be included in an ontology (for example, the term *results*) and those that seem like potential MeSH candidates such as *placebo values*.

While this mapping did not restrict on the semantic category, and while the results might have been different if we had restricted on the semantic category, the results obtained indicate a gap that would have occurred as the result of trying to map the text to an ontology before identifying integral parts of a comparative sentence. The authors use a range of terms when describing the entities being compared and this range would be very difficult to specify in advance. The surface forms of entities in the text are frequently not at the right level of granularity—sometimes they can be too specific and sometimes too general in nature—to be easily mapped against an ontology. The need to predefine entities would necessarily result in the loss of some information which is why the methods described in this proposal do not predefine concepts/entities and entities emerge through the grammatical structure of the sentence.

And yet, as it will be argued in Chapter 3 of the dissertation, semantic classes of nouns in an ontology can be helpful when trying to predict the class of the noun. Inferring the semantic class of the noun by focusing on the head noun of the compound noun increased the precision and recall with which entity 1 and endpoint were identified. However, limiting the nouns to only one or several semantic category would most likely end up being a limitation. The main advantage of the method described in this dissertation is that it does not restrict by a semantic category and the categories emerge through the syntactic structure of the sentence.

One of the conclusions of the pilot study was that while the levels of precision and recall for each of the three facets are not unsatisfactory and can, taken individually, reveal interesting trends, the automatic extraction of comparative facets would benefit from a model that boasts a higher precision and recall for the endpoint and entity 1 identification. This section outlines the steps and heuristics that were considered to improve the model for these two facets. Also, the following section explains how the method of identifying comparison sentences was altered with the aim of achieving better precision with respect to identifying direct comparative sentences in scholarly articles.

CHAPTER 3: DISTRIBUTION OF COMPARISON SENTENCES AND THEIR FACETS

3.1 DISTRIBUTION OF COMPARISON SENTENCES ACROSS DIFFERENT ARTICLE SECTIONS

Pilot study established that the classifiers would benefit from higher precision and recall that they identify and separate the roles of the first compared entity and the endpoint in a comparative sentence. Pilot study also established that not all candidate comparison sentences retrieved can be considered *direct* comparative sentences. This chapter is dedicated to improving the precision with which classifiers identify the roles of the first compared entity and endpoint and also the precision with which direct comparison sentences are extracted from a collection. For this, comparison sentence distribution in the articles as well as individual comparison facet distribution is analyzed. First a few words about comparison sentence distribution.

As earlier chapters emphasized, the focus of this thesis is on *direct* comparative sentences. To be considered a direct comparative the sentence needs to contain a minimum of two compared entities, an indication of the basis on which the entities were compared (endpoint), and also the result of comparison or the indication of whether a property in an entity *increased*, *decreased*, was the *same* or *different* compared to another entity. And yet, not all the candidate comparative sentences are direct comparatives. Many retrieved candidate comparative sentences do not contain one of the required facets (e.g. refer to entity 1 or entity 2 through an anaphoric reference) or simply do not communicate the result of comparison or the semantic relation that ties the entities and the endpoint. To increase the precision with which *direct* comparative sentences are identified using the lexico-syntactic patterns and heuristics as outlined in the pilot study (Chapter 2), the comparison sentences were examined with respect to the section of the article they appear in the three TREC Genomics journals used in this dissertation. The main question that guided this analysis was: can precision of detecting direct comparison sentences be improved if we focus only on particular section/s of the article?

The pilot study considered the comparison sentences throughout the entire article regardless of the section in which they appear. And yet, it is likely that comparison sentences tend to occur more often in certain sections than in the others. Also, it is possible that the information contained in one comparison sentence in the article—for example, in the abstract—is repeated later on in the article, for example, in the Results or Discussion section. Earlier work in argumentative zoning (Teufel & Moens, 1999) showed that different types of argumentation occur across different sections of the document and are not restricted to only one section. Teufel & Moens (1999) demonstrated that although it is possible to restrict the annotation area to Introduction, Abstract, and Conclusion because these areas might contain clearer arguments, these sections, in general, contain a number of different types of arguments and are not restricted to one argument only. This study focuses on one particular argument, a direct comparison which, partially, although not fully, corresponds to Teufel's Contrast category. Because the overlap with the Contrast category is only partial, relying on the information about the distribution of this particular argument across different collections is not sufficient. The first step involved an analysis of the overall number of comparative sentences across articles in *Diabetes* collection. Figure 9 indicates that the overall number of comparative sentences in *Diabetes* varies. Interestingly, only 21 articles from TREC *Diabetes* journal collection do not contain at least one comparison sentence. 2,121 out of 2,141 articles or 99% contain at least one candidate comparative sentence. The number of comparative sentences across articles ranges from a minimum of 1 to a maximum of 49 with mean of 10 and median of 9 sentences.

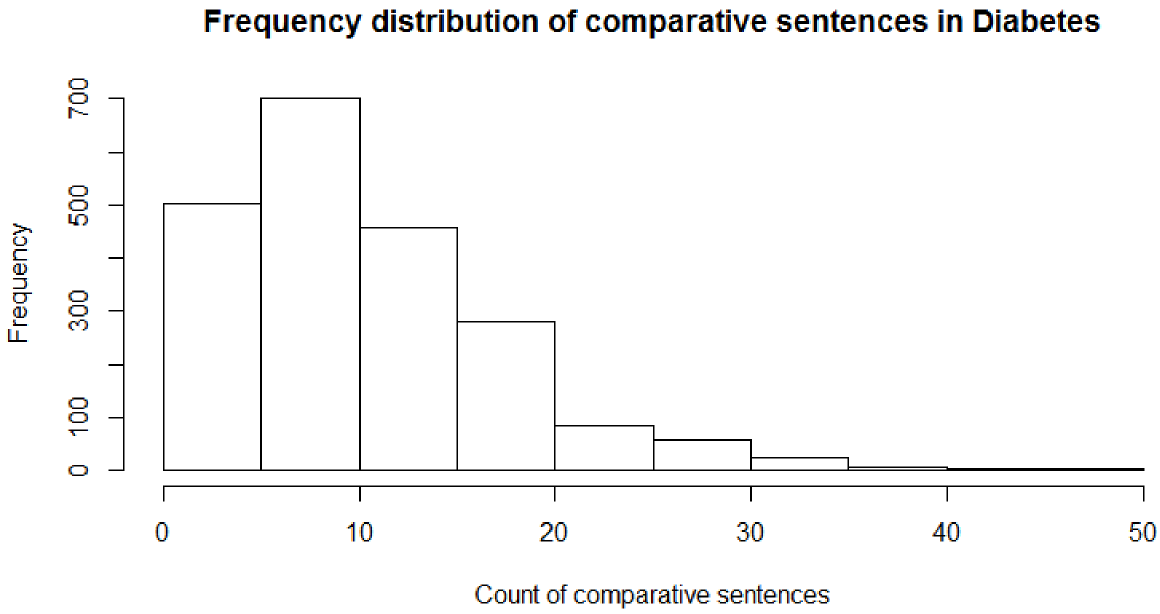

Figure 9: Frequency distribution of comparison sentences in *Diabetes* journal

This result poses some interesting questions, such as, what is the nature of the articles that do not include any comparison sentence? How are they different or similar to the other articles in the collection that contain at least one comparison sentence? What is the nature of the studies that include many comparison sentences? Is there a threshold number of comparison sentences in an article that indicates a comparative study? More importantly, to what extent can comparative sentences in an individual article be used to summarize the results of a particular article? A closer analysis of these questions is left for future work.

To evaluate, however, where in the article direct comparison sentences are more likely to appear, the comparative sentences need to be mapped to the section they originated from, i.e., Abstract, Introduction, Result, Discussion, or Conclusion, if they indeed come from a particular section. Relying on the set of heuristics that takes into account different heading variants, the following mapping process was devised.

Section name	**Mapped to**	**Resulting section**
Abstract	➔	Abstract
Introduction; Background	➔	Introduction
Research Design and Methods; Materials and Methods, Methods	➔	Method
Preliminary Results; Early Results; Results; Result	➔	Result
Conclusion; Summary	➔	Conclusion

Table 13: Mapping of section names in scholarly articles

Although section name variants in Table 13 may not be exhaustive and may not include every variation under which a particular section name appears, this mapping strategy provided a relatively good coverage for a large number of articles in the TREC Genomics collection.

Table 14 indicates the results of the mapping process on the three TREC 2006 Genomics journals.

Section	Diabetes	Carcinogenesis	Endocrinology
ABSTRACT	100.00%	100.00%	99.84%
INTRODUCTION	41.80%	98.62%	98.61%
METHOD	97.34%	88.56%	98.28%
RESULT	89.21%	88.92%	98.36%
DISCUSSION	89.35%	85.96%	97.25%
CONCLUSION	1.21%	4.24%	2.06%

Table 14: Presence of journal sections in three TREC 2006 Genomics journals

As Table 14 indicates most of articles in the three journals contain an Abstract. As a matter of fact, some of the articles, such as, for example, short communications, only contain an Abstract. All of the articles in *Diabetes* and *Carcinogenesis* journals that are part of TREC Genomics collection contained an abstract. In *Endocrinology*, 99.84% of the articles contained an Abstract. Introduction is not very frequent in *Diabetes* (41.80%) journal but is much more common in *Carcinogenesis* (98.62%) and *Endocrinology* (98.62%). Method is a common section across the three journals and occurs approximately 98% of the time in *Diabetes* and *Endocrinology* and 88% of the time in *Carcinogenesis*. Result is another frequent section that occurs in close to 90% of the articles in *Diabetes* and *Carcinogenesis* and in 98% of the articles in *Endocrinology*. *Discussion* occurs in close to 90% of the articles in *Diabetes*, close to 86% of the articles in *Carcinogenesis* and 97% of the articles in *Endocrinology*. Conclusion is the least common section across the three journals. The results indicate that *Diabetes* and *Endocrinology* articles contain the Conclusion section in only 1-2% of the articles and 4% in *Carcinogenesis*.

It is possible, of course, that some of the articles contain the more common section names (Abstract, Discussion, Result) but the variant of the section name used in the article was not recognized during the mapping process. In general, however, the results indicate that the majority of the articles in the three journals used in this study follow a more traditional and structured format of writing scholarly articles that indicate an Abstract, Method, Result, and/or Discussion section. Conclusion section is the only exception. The overall result, however, is that the three journals follow a more conventional structure of the article because of the presence of these major sections. In *Diabetes* journal, this convention is followed by 1,933 out of 2,041 articles, or 90%, in *Carcinogenesis*, 1,790 out of 1,958 articles, or 91%, and in *Endocrinology*, 5,000 out of 5,050 articles, or 99%, follow such structure because they

To return to the original question of where in the article comparative sentences tend to occur, a subset of 20 articles from *Diabetes* journal that were used in the Metformin case study (see Section 2.9) was utilized. The total number of identified comparative sentences in this sample of articles was 221. Analogous to detecting direct comparison sentences in the pilot study, the sentences that included anaphoric reference were considered direct comparison sentences. 194 out of 221 sentences satisfied these criteria resulting in 0.88 precision. This represents a spike of 0.05 from earlier analysis where 1,000 randomly selected sentences were extracted and analyzed. The slight spike in precision may be due to the nature of the articles that involve drug comparison which, in turn, may be more likely to contain direct comparison sentences. The section names for candidate comparison sentences from 20 articles were then identified using the mapping process described earlier and their distribution across different section names analyzed.

As Figure 10 indicates, comparison sentences in this sample are more frequent in the Result and Discussion section than in Method, Introduction and Conclusion. Comparison sentences occur relatively frequently also in the Abstract section.

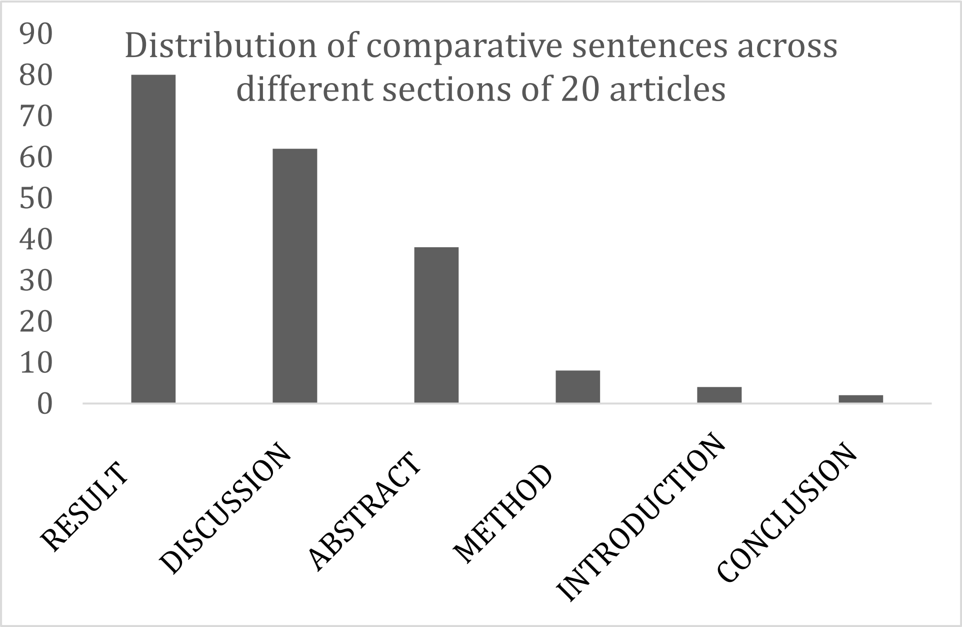

Figure 10: Distribution of comparison sentences across different sections of 20 articles

The same analysis was repeated on the entire TREC *Diabetes* journal collection.

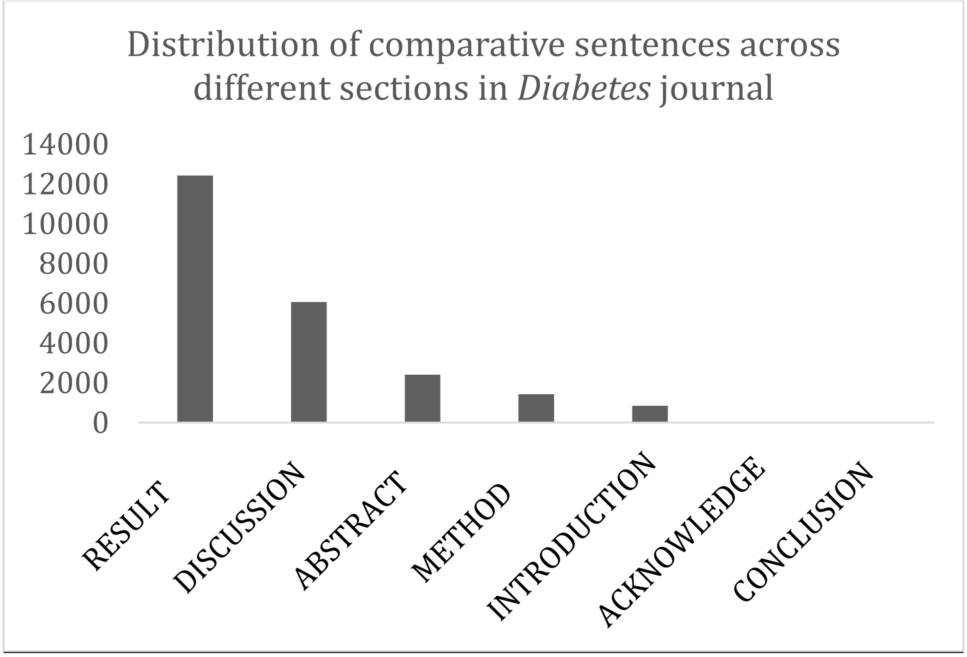

Figure 11: Distribution of comparison sentences across different sections in *Diabetes* journal (23,034 sentences)

Figure 11 bar chart closely resembles the bar chart of Figure 10. As Figure 11 indicates, the section that has the most comparative sentences is the Result section which is then followed by Discussion and Abstract. Surprisingly, Conclusion section contains only 8 comparative sentences. This seems to indicate that comparative sentences do not feature prominently in the Conclusion of the articles in *Diabetes* journal. However, earlier analysis revealed that a majority of articles in *Diabetes* do not contain a Conclusion section. Frequently, the articles end with a Discussion section which explains low frequency of comparative sentence in the Conclusion section.

This distribution analysis of comparative sentences indicated that Results, Discussion, and Abstract sections are the most relevant sections for identification of comparative sentences. We are clearly not interested in the Acknowledgement section as the comparisons expressed there would not likely communicate the results of the study under consideration.

Although, as Table 14 indicated, Abstract is a common section across the three journals, the articles that contain only an Abstract section were not considered. The reason for this is that direct comparison sentences that report results of the experiments are of primary interest in this study and such sentences are more likely to be reported in the articles that follow a more conventional structure of the article and contain either the Method, Result or Discussion section in addition to Abstract. As mentioned earlier, in *Diabetes* journal, this convention is followed by 1,933 out of 2,041 articles, or 90%, in *Carcinogenesis*, 1,790 out of 1,958 articles, or 91%, and in *Endocrinology*, 5,000 out of 5,050 articles, or 99%, follow such structure.

Given that direct comparison sentences are the main focus of this study and given that comparative sentences in the biomedical science articles tend to occur more in the sections such as Results, Discussion, Abstract than the others, a random sample of 1,000 comparison sentences across all sections of the articles was extracted for an additional analysis. Unlike in the Pilot study (Chapter 2) which used 65 comparison anchors, this extraction of candidate comparison sentences used a revised and expanded set of 70 comparison anchors (see section 4.1). Figure 12 indicates the distribution of 1,000 comparison sentences across different sections of the article:

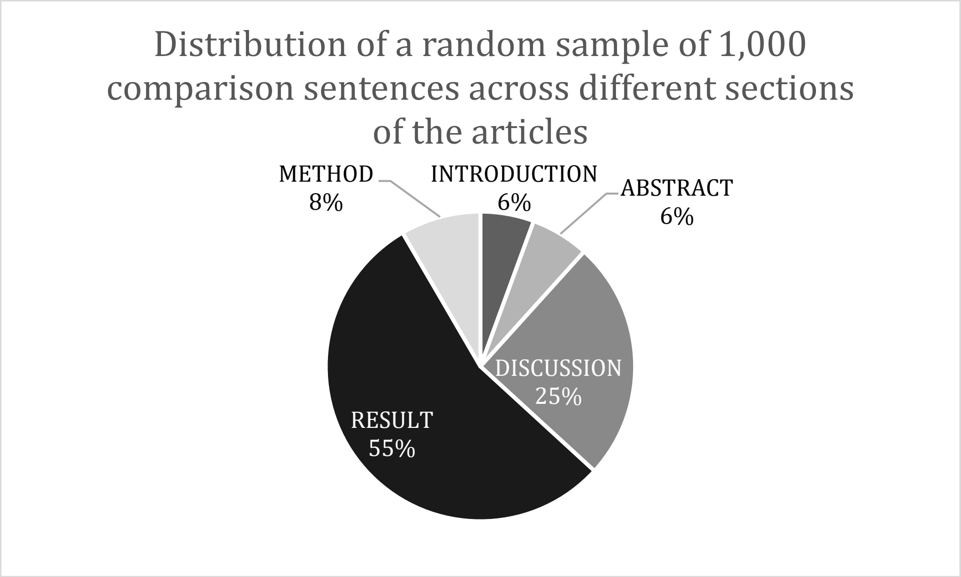

Figure 12: Distribution of a random sample of 1,000 comparison sentences across different sections of the articles

Similar to Figure 10 and 11, Figure 12 indicates that the majority of candidate comparison sentences come from Result, Discussion. Approximately 8% of the sentences retrieved come from Method and 6% from Introduction.

3.2 TYPES OF COMPARISONS

Similar to the Pilot study (Chapter 2), the analysis that follows aimed to establish the percentage of comparison sentences in the sample of 1,000 that were direct comparisons. However, unlike in the Pilot study, it was decided to take into consideration the meaning and information that the sentences that do not fit the definition of a direct comparison sentence convey and assign them labels based on their different qualities, such as their meaning, clarity and complexity. This analysis, however, does not aim to analyze the comparison sentences with respect to whether they are gradable or non-gradable. The analysis conducted here is interested in establishing whether the sentences selected can be deemed a direct comparison sentence and if not, identify the reasons why it cannot be deemed a direct comparison sentence. Inclusive of the direct comparison label, this analysis revealed nine different groups of sentences that were all collectively retrieved as comparison sentences. The labels assigned to each comparison sentence and their meaning are included in Table 15.

Label	**Meaning**
Direct comparison	The sentence contains two or more entities that are compared on two or more aspects/endpoints. The sentence also communicates the result of comparison.
Anaphoric reference	At least one of the comparison facets is only implied and not mentioned directly.
Comparison statement	Only a statement that a comparison was made between two or more entities. Typically, the sentence does not communicate the result of comparison.
Unclear comparison	The sentence is unclear. It is difficult to intuit the exact meaning of the sentence without considering the surrounding context.
Method	The comparison sentence describes the method of the experimental study rather than the result.
Complex comparison	In addition to a comparison statement the sentence communication additional information and claims. The comparison claim inside a complex sentence is usually either nested or conjunctive.
Explicit statement	The comparison anchor in the sentence is used to express an explicit statement rather than a direct comparison.
Speculative comparison	Comparison sentence that communicates a speculative result.
Special case comparison	Rather than having two or more entities that are being compared on a particular aspect, the sentence compares two aspects with respect to one entity.

Table 15: Different types of comparison sentences

A few words should be mentioned about different labels assigned to different types of comparison sentences.

Direct comparison – the main focus of this study, references the minimum of two compared entities, the endpoint and also the result of comparison.

Anaphoric reference – a comparison where one of the required facets is not mentioned directly and needs to be inferred from the preceding context. Although a comparison, this particular type was considered a separate type from direct comparison. This thesis does not propose a mechanism for inferring the required facet from the surrounding context which is left for future work.

Comparison statement – a type of sentence that only states that a comparison between two or more entities took place but it does not state the result of comparison. This sentence can also include the information about the endpoint of comparison. Because this thesis is only interested in sentences that include each of the four facets, this type of sentence is assigned its own category.

Unclear – a type of comparison that is difficult to understand and untangle from the surrounding content. This type is typically inserted inside a longer, complex sentence. Although this thesis currently does not offer mechanism for automatically filtering such sentences, this label indicates that the complexity of comparative sentences is something that future work will need to grapple with.

Method – a type of comparative sentences that typically describes the method of the study which semantically sets them apart from comparative sentences that indicate the result of the study.

Complex comparison – a type of comparison that indicates nested and conjunctive comparisons inside a single sentence that express multiple comparative claims.

Speculative comparisons – a type of comparison that is speculative in nature. The current mechanism does not filter such sentences out but recognizing the difference between sentences that indicate that something may have changed rather than expressing that something has changed represents an important difference in meaning that future work will need to pay attention to.

Explicit statement – a type of sentence in which comparison marker such as *similar to* or *different from* rather than expressing a comparative relation expresses an explicit statement.

Special case – a type of comparison that does not follow the Claim framework model of a comparative sentence (Blake, 2010).

These labels are not mutually exclusively and obviously there is some amount of overlap between them. For example, it is not always clear which sentence should be labeled Complex versus Unclear. A speculative comparative sentence can also contain an anaphoric reference. Labels used in Table 15 represent a way to separate direct comparison sentences that are the main focus of this study from the rest of comparative sentences that are commonly found in the collection of biomedical scholarly articles.

Each of the sentences in the sample of 1,000 was assigned one of the nine categories. Of particular interest is the nature of the sentences that occur in the Method section. 84 (approximately 8%) sentences come from the Method section (see Figure 12) whereas the method of annotation categorized 62 sentences as having originated in the Method section. The number of sentences that was annotated as having originated in the Method section and those that actually came from the Method section was 40 which represents 0.74 agreement. Interestingly, 22 sentences which appeared on the surface to have originated in the Method section actually originated from the Result or Discussion or Conclusion section. A direct comparison label was assigned to 9 sentences and Comparison statement label to 24 sentences (only a statement that a comparison was made) and yet, upon checking the section they originated from, it was established that these sentences came from the Method section. As a matter of fact, comparison statements (only a statement that a comparison was made) frequently originate in the Method section of the article. Here, however, is an example of the sentence that comes from the Method section that was indeed annotated as the Method comparison sentence:

(27) Hepatic glucose release was calculated as the difference between the systemic glucose release and renal glucose release. 12765948

Rather than communicating the result of comparison, this sentence sets the stage for communicating the result of the study. These results indicate that Comparison statement and Method comparison are good candidates for collapsing into one. When combined, Comparison statement and Method categories make up a total of 64 out of 84 sentences that expressed comparative relations that were not of interest in this study. What complicates things, however, is that 18 out of 62 sentences were annotated as having originated in the Method section but they actually came from the Result section. The following is an example of such sentence:

(28) Hyperlipidemia was present in 50% and hypertension in 70% of the patients, without differences between the groups. 15448095

9 out of 84 sentences that came from the Method section were annotated as direct comparison sentences and they, in fact, are direct comparison sentences. The following sentence serves as an example:

(29) The cumulative incidence of diabetes in BB.7_b animals is similar to that observed in BBDP rats originating from the BRM colony (data not shown). 12351436

What the results so far indicate is that a direct comparison sentence that is of primary interest in this study is not restricted to a particular section or particular sections of the article. Direct comparison sentences that express the result of comparison are, sometimes, included in the Method section as well. And yet, in general, the Method section tends to contain statements that a comparison was made and thus sets a stage for the results to be communicated later on in the article.

Although, obviously, the division between Method, Result and Discussion sections is not clear cut and although direct comparisons that communicate the result do occur in the Method section, the large majority of sentences in the Method section (76%) are not of interest in this study. The sentences that occur in the Method section, at least in the collection that is of interest in this study (a subset of TREC Genomics collection), communicate, for the most part, the conditions of the experiment. Additionally, Method section, in general, is more likely to contain a mention that a comparison was made rather than the result of the comparison. For this reason, it was decided that comparison sentences retrieved from the Method section will not be included in the main experiment.

Another candidate for exclusion was the Introduction section. As Figure 12 indicates, only 6% of randomly extracted comparison sentences originated in the Introduction section. Although earlier work (Teuefel & Moens, 1999; Blake, 2010) indicated that relevant claims and particular types of arguments can be found dispersed throughout an article rather than concentrated within one section of the article, when we are focusing on one particular structure, such as, for example, direct comparison, it is reasonable to assume that direct comparison expressed in the Introduction or Background section may likely refer to prior work and prior knowledge rather than to the current results of the experiments. For example, consider the following sentence:

(30) It has been recognized for >15 years that methylation patterns in tumor cells are

altered relative to those of normal cells. 10688866

Although this sentence communicates the result of a comparison and although it can be considered a direct comparison, this sentence primarily establishes prior knowledge that for more than 15 years methylation patterns in tumor cells have been altered relative to those of normal cells. The following is another example of the sentence that occurs in the Introduction or Background section and refers to prior knowledge:

(31) There are considerable data to support the concept that the type of fat or fiber is actually more important to tumor development than is the amount of either of these components in the diet. 10910952

While the information communicated through such sentences is certainly noteworthy, we may proceed with an assumption that it is more likely that a comparison sentence in the Introduction section refers to the prior knowledge and background information rather than to the results of the current study. Given that this study is interested in the potential of comparison sentences to summarize the result of the experiments and given that the inclusion of the current result can be deemed more pertinent for inclusion in the summary than the earlier result, comparison sentences retrieved from the Introduction section were excluded from the main experiment. As mentioned earlier, so were the comparative sentences extracted from the Method section. The comparative sentences in the random sample that came from the Method and Introduction sections were then replaced with randomly extracted sentences from either the Abstract, Result, Discussion or Conclusion.

Figure 13 indicates how different types of comparison sentences in the revised random sample of 1,000 candidate comparative sentences are distributed. These sentences now only come from the Abstract, Result, Discussion and Conclusion sections.

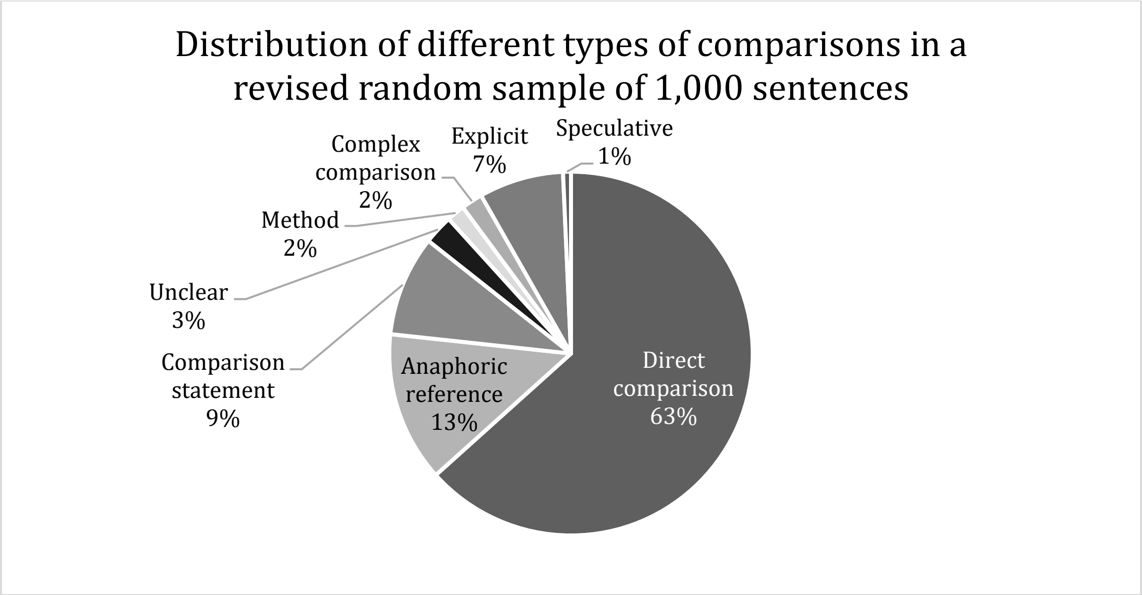

Figure 13: Distribution of different types of comparison sentence in a revised random sample

As Figure 13 indicates, the largest percentage of randomly selected sentences, 63% of them, are direct comparison meaning that they include the references to each of the facets (two entities and the endpoint) and also communicate the result of the comparison and the change that has occurred. 13% of the sentences include an anaphoric reference meaning that in those sentences at least one of the facets is not referred to directly but can be inferred from the surrounding context. Both of these categories were considered to be direct comparisons in the pilot study. According to Figure 13, two largest categories of comparison sentences retrieved that do not satisfy the criteria of direct comparison sentence definition are Comparison and Explicit statements. These two categories are related in the sense that each uses a comparison anchor that rather than communicating the result of the comparison either indicates that a comparison has taken place (comparison statement) or helps form an explicit statement.

Here is how the distribution of these categories looks like when mapped against the article sections they come from.

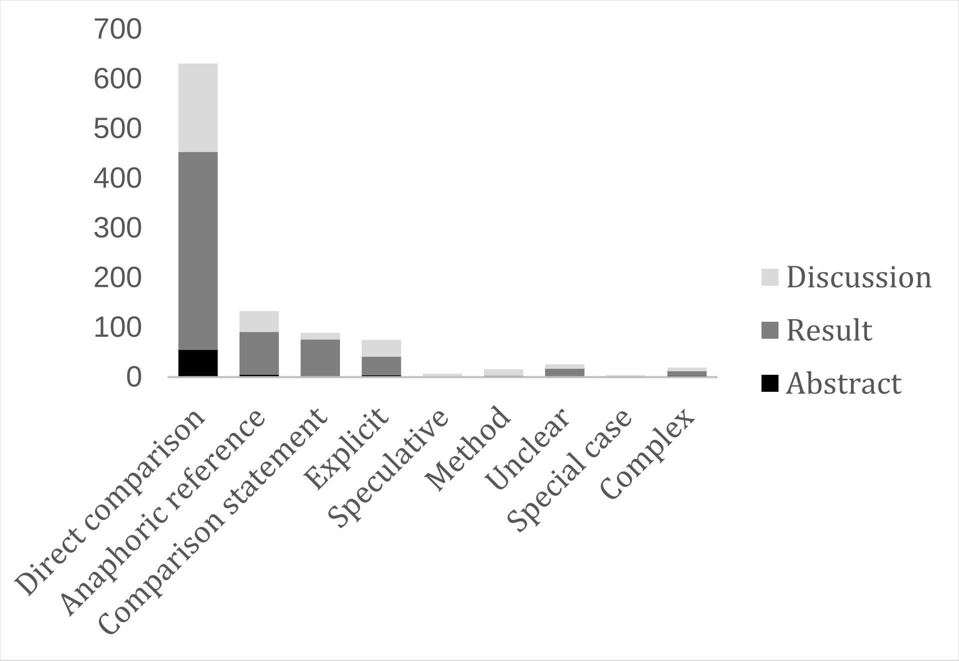

Figure 14: Distribution of different types of comparisons across three article sections

Figure 14 indicates that direct comparison is by far the largest category and also that each of the specified categories can be found not only in one section of the article but across different sections.

A few words should be said about the categories such as Unclear (3%), Complex (2%), Speculative (1%), Special case (only four instances). Although comparison sentences that received these labels were not considered direct comparison, it is not clear that each of these categories does not, in fact, contain a valid direct comparison sentence. Consider, for example, the following sentence that was annotated as an unclear comparison:

(32) Note that the amount of 35_S-labeling was similar in all lanes, whereas 125_I-iodine incorporation was increased with LPO. 16037381

Although there are certainly comparisons buried in this sentence, the true meaning of the sentence is easier to comprehend if we consider the larger context of the sentence. For sentences that are marked as Unclear it is often difficult to establish what the compared entities in the sentence are versus the endpoint that they are being compared on. Additionally, such sentences frequently resemble an explicit statement rather than a direct comparison.

Speculative comparison statements, as the label indicates, convey a speculative comparison sentence: Consider the following example:

(33) Because of the differences between inulin and insulin itself, whether delivery of the bioactive hormone is increased remains speculative. 15504953

Rather than communicating the change that has occurred, this sentence communicates that it is *not certain* whether the delivery of the bioactive hormone is increased because of the differences that exist between inulin and insulin. Put differently, there is a possibility that the delivery of the bioactive hormone is increased but the finding remains inconclusive. The following is another example of a speculative comparison sentence:

(34) To what extent the cell dispersion procedure promotes a biological situation similar to inflammation or injury is unclear. 9348205

Although cell dispersion procedure and inflammation or injury are compared in this sentence with respect to biological situation, the extent of this promotion remains unclear. Sentences such as this one contain a comparison that is cloaked in speculative language. While this dissertation does not aim to recognize or retrieve this particular type of comparison sentence nor does it propose a strategy for eliminating this particular type of comparison sentence, one potential strategy for eliminating speculative comparisons would consist of detecting hedge terms in comparison sentences.

Another small category in the pool of retrieved comparison sentences is the, so-called, *special case* category. The sentences assigned this label typically do not follow the claim framework definition of a comparison that implies a comparison of two or more entities on one or more aspects. The following is an example of this category:

(35) Conversely, at higher (nanomolar range) concentrations the effects of melatonin resulted in a stimulation of GH secretion, in marked contrast with the cAMP levels, which continued to decrease. 12960030

In this sentence, the effects of melatonin at higher concentrations are analyzed with respect to its effect on the stimulation of GH secretion as well as on the cAMP levels. While this sentence represents a comparison sentence and while it is recognized as a comparison sentence, it does not follow the logic of a direct comparison sentence as defined through Claim Framework. Rather than containing two entities that are being compared, this sentence expresses the comparison of two endpoints (*GH secretion* and *cAMP levels*) with respect to another entity or, better to say, another endpoint: *higher concentrations of melatonin*. In sentences such as this one, it appears as though the entities and endpoints have switched their roles which represents another reason why comparison sentences are so inherently complex.

Finally, to understand the Complex comparison sentence, we need to rely on the surrounding context which puts them in close proximity to Anaphoric references group. In some cases, though, the sentences that were annotated as Complex combine a Special case category and a direct comparison as in the following example:

(36) However, FM1-43 was reported to more likely stain the lipid membrane by unrestricted lateral diffusion through the membrane than by aqueous diffusion , and therefore, the kinetics of the rising phase of FM1-43 fluorescence might be similar between these exocytic processes. 16123364

Among other things, this sentence contains a direct comparison: *FM1-43 fluorescence may be similar between unrestricted lateral diffusion and aqueous diffusion* and yet establishing the noun role in a sentence such as this one represents a more complex task than in a more representative direct comparison sentence. This sentence could also be categorized as a speculative comparison sentence because of the presence of the verb *might*. This sentence also contains an anaphoric reference based on which it can be assigned Anaphoric category.

These four groups of comparison sentences account for approximately 8% of the randomly extracted comparison sentences. The real challenge, for future work, however, will be the separation of sentences that communicate the result of a comparison from those that merely express that a comparison was made (in the current random sample, 9%) or that express an explicit statement (in the current sample, 7%). These two groups combined (16%) represent the largest obstacle that stands in the way of retrieving direct comparison sentences with higher precision. Consider the following sentence:

(37) In men, a single injection of a relatively large dose of rhFSH (3000 IU) resulted in less than a 2-fold increase in inhibin B levels. 12639898

In this sentence, the comparison anchor *less than* communicates the *result* of an injection of a relatively large dose of rhFSH rather than compares the result of this action to another.

Although the elimination of sentences form the Method section has reduced the number of sentences that only communicate that a comparison was made, it has not completely eliminated them. In fact, comparison sentences from the Result section, as the following two sentences will indicate, can, on occasion, communicate information that is more commonly expected in the Method section. Consider the following two sentences:

(38) Differences between groups were analyzed by the Student's t test. 12663468

(39) Total RNA from two control mice (nos 2 and 3) with weight gain similar to the DCA-treated mice were used as controls. 11470764

The sentences such as this one typically provide the background information about the experiment and as such do not convey the result of the comparison.

The revised set of 1,000 sentences was then analyzed with respect to the article section they come from.

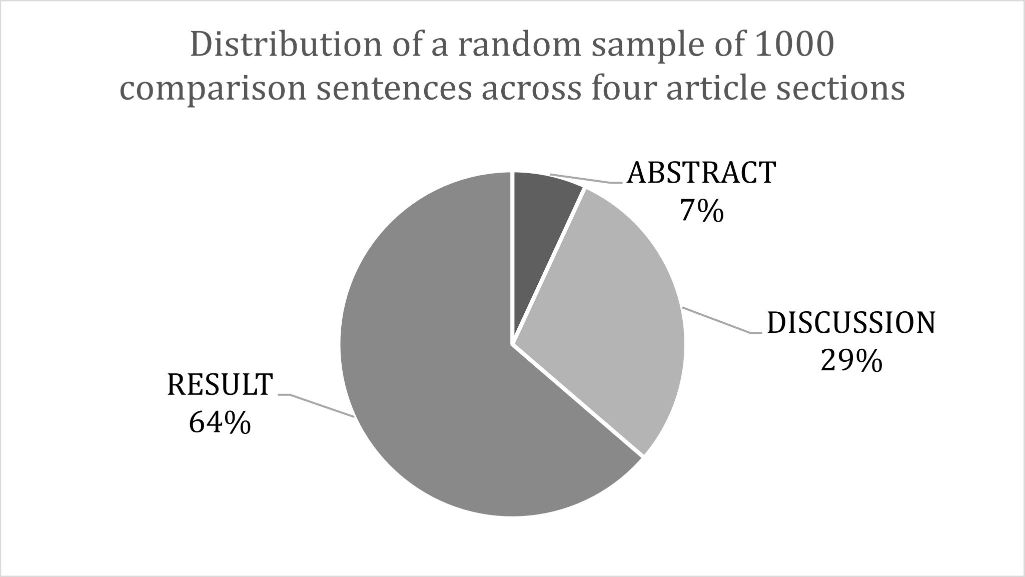

Figure 15: Distribution of a *revised* random sample of 1,000 comparison sentences

As Figure 15 indicates, the majority of sentences (64%) came from the Result section. This is followed by Discussion (29%) and Abstract (7%). None of the randomly extracted sentences came from the Conclusion section (this is a relatively small group of sentences in each of the three TREC journals). For comparison purposes, Figure 16 indicates a distribution of comparison sentences across the three sections in three journals: *Diabetes*, *Endocrinology* and *Carcinogenesis*:

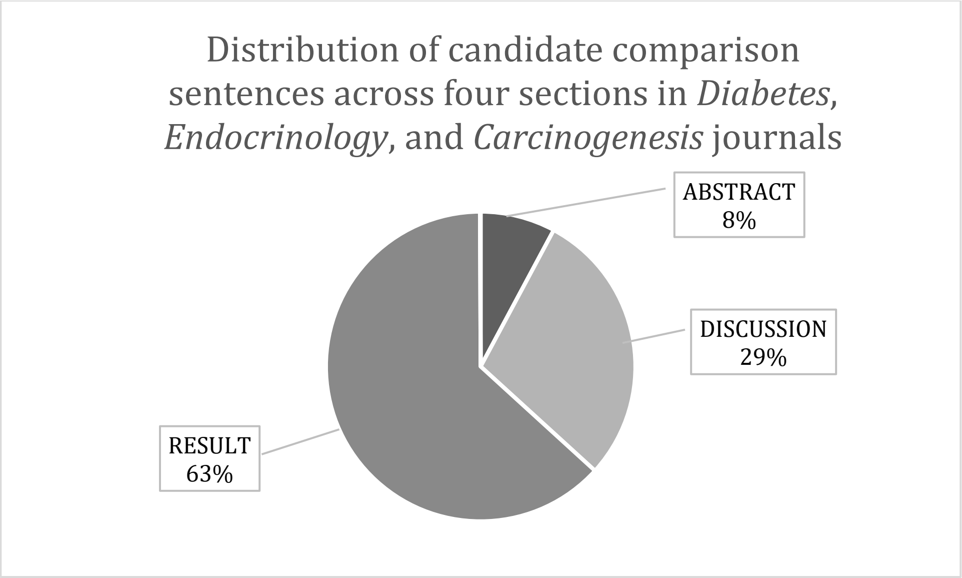

Figure 16: Distribution of candidate comparison sentences across three sections

Analogous to earlier result (Figures 10, 11 and 12), the largest number of comparison sentences is extracted from the Result section, then Discussion which is followed by the Abstract section. Conclusion section is rare in this particular collection, only 64 sentences were identified as having originated in the Conclusion section. Conclusion section is not represented in the pie chart.

Given that the focus of the analysis now is only on particular article sections, it seems as though the number of candidate comparison sentences may drop. The following chapter will demonstrate that the number of candidate comparison sentences indeed did drop in *Carcinogenesis* and *Endocrinology* journals but, interestingly, it increased in *Diabetes*.

3.3 DISTRIBUTION OF COMPARED ENTITIES AND ENDPOINTS ACROSS DIFFERENT ARTICLE SECTIONS

The earlier sections in this chapter analyzed the distribution of comparative sentences across different sections of the article. Another distribution of interest is that of comparison facets throughout the article. Of primary interest is the question: Are compared entities more likely to appear more frequently in certain sections of the article?

Also of interest are the questions of whether endpoints are more likely to be discussed in the Method, Results, or Discussion section and how does the distribution of compared entities differ from the distribution of the endpoint throughout an article?

Figure 17 below indicates that there is little difference in the overall distribution of compared entities and endpoints across 20 *Diabetes* journal articles. Similar to comparative sentences, the largest number of compound nouns and endpoints appear in the Result section followed by the Discussion section.

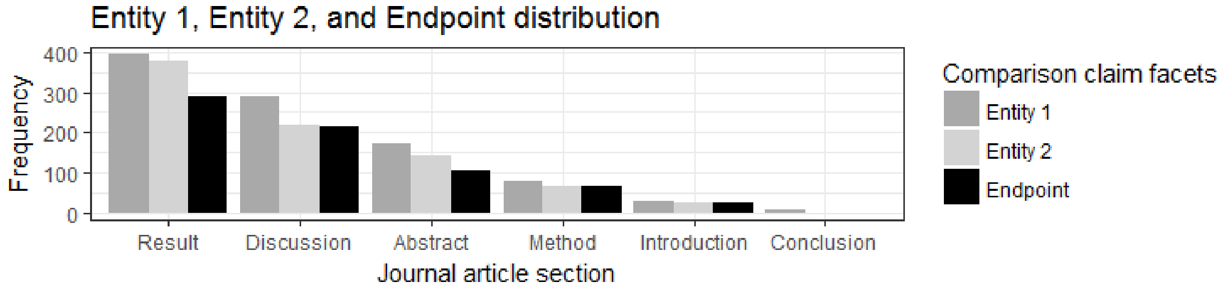

Figure 17: Entity 1, Entity 2, and Endpoint distribution across different sections of 20 articles

This analysis also established that the overall frequency of a particular compound noun may not be a useful indicator of the role that the compound noun plays in a comparative sentence. Although in most articles the compared entities occur with higher frequency than the endpoints this was not true in all of the 20 examined articles. In several articles, the noun that was identified as the endpoint was the noun phrase that occurred with highest frequency in the article. Depending on the topic of the article, the focus may be on endpoint rather than on a compared entity.

3.4 NATURE OF ENDPOINTS IN BIOMEDICAL ARTICLES[4]

Given that the overall distribution of compared entities did not seem to provide a useful and reliable method for establishing the nature of the noun in the sentence, the focus has shifted to the nature of the verbs that endpoints and compared entities are associated with in comparison sentences.

Within the context of biomedical scholarly articles, endpoints frequently indicate a substance or a property that has either undergone some change (gradable comparison), has stayed

[4] The text of this subsection also appears in the following publication:

Lucic, Ana & Blake, Catherine. (2016). Improving Endpoint Detection to Support Automated Systematic Reviews. AMIA conference, November 12-16, 2016. Chicago, IL.

the same (non-gradable comparison) or is simply similar or different to another entity that shares this property (non-gradable comparison). The analysis of the most frequent head noun of endpoints in the collection revealed *level* or *concentration* as a very frequent head noun of the endpoints. Common to both *level* and *concentration* is that they can be quantified. Given that, in general, the Method section in the collection of biomedical articles provides the information about the laboratory and study procedures it was decided to establish the nature or class of verbs that are associated with endpoints versus compared entities in the Method section. The verbs associated with the endpoints versus compared entities were thus compared in the subset of 20 articles that feature metformin as a compared entity in a comparison sentence.

Verb	**Compared entity**	**Verb**	**Endpoint**
admitted	subjects	assessed	development
discontinued	troglitazone	calculated	mass
dissolved	metformin	computed	amyloid
infused	subjects	determined	mass
stopped	troglitazone	estimated	sensitivity
titrated	treatment	maintained	HMEC-1
treated	subjects	quantified	proportion
purified	AMPPK1	triggered	opening

Table 16: Compared entity and endpoint associated verbs in the Method section

Table 16 indicates that the verbs that indicate measurement and quantification such as *calculated*, *computed, quantified* are associated with endpoints in the Method section. The verbs such as *admitted, discontinued, treated, used* are mostly associated with compared entities. This analysis revealed that the endpoints are frequently nouns that can be measured or quantified. Indeed, the nouns such as *mass*, *sensitivity*, *proportion*, *opening*, *development* belong to the class of uncountable nouns that lend themselves to measurement and quantification. Although *troglitazone* and *metformin* also lend themselves to measurement in the sense that we can administer different levels and measurements of these two drugs they do not exactly belong to the same class of words such as *mass*, *proportion* and, *concentration*. In the UMLS semantic network hierarchy, *mass*, *proportion*, and *concentration* are each associated with the semantic class Quantitative concept whereas it is not easy to establish this connection for the drugs *troglitazone* and *metformin*. Interestingly, though, *level* rather than being identified as a Quantitative concept was identified as a Qualitative concept in the UMLS semantic network hierarchy. This analysis prompted me to establish to which degree the association of the noun with its semantic class—in the case of the endpoint role with the Quantitative concept—can help us separate the endpoint from the compared entity. The hypothesis is that the addition of a semantic class to the feature set would help separate the first compared entity from endpoint. The question, however, was how to obtain the information about the semantic class for compared entities and endpoints? From an ontological point of view, entities that occur in comparison sentences can be matched to an ontological class such as species or population group. For example, the Unified Medical Language System Metathesaurus semantic class Population group can be seen as helpful for their identification. Consider sentence 6 that was also referenced in Chapter 1 (section 1.1) of this dissertation:

(6) The plasma insulin concentration [endpoint_A] at 8 weeks [endpoint_A] of age [endpoint_A] and the pancreatic insulin content [endpoint_B] and the beta-cell mass [endpoint_C] on day 8 [endpoint_BC] and 8 weeks [endpoint_BC] of age [endpoint_BC] in STZ-treated rats [entity 1] were *severely* [relation modifier] *reduced* [relation] compared with those of normal rats [entity 2] (P < 0.001). 14988244

In this sentence, *STZ-treated rats* and *normal rats* represent two entities that can be seen as two population groups that were compared. And yet matching these two concepts to their semantic classes using the UMLS Metathesaurus does not bring us to the Population but rather to the Animals class. Neither *STZ-treated rats* nor normal rats has its full match in the UMLS because they represent very specific groups of rats: rats treated with streptozotocin versus normal rats. With both of these examples, however, it is the head noun—rats—that provides sufficient basis for inferring the semantic class for these two entities—Mammals—and then, through its parent relation, to Vertebrae and Animals higher up in the hierarchy. Animals, however, do not link directly to Population group in the UMLS. In the higher levels of semantic network, Animals is an Entity, the broad type used for grouping conceptual and physical entities. Population group in the UMLS Metathesaurus is defined as “an individual or individuals classified according to their sex, racial origin, religion, common place of living, financial or social status, or some other cultural or behavioral attribute” and as such does not extend to Animals. While we can decide to treat all the nouns that match to the Animals semantic class as Population group, within the context of comparison sentences in biomedical literature, there is also a danger in extending and widening the semantic pool too far because it may result in many false positives.

Similarly, matching endpoints to an ontological class is far from being a straightforward process. As it was demonstrated earlier, although *property*, *mass* and *concentration* are associated with the Semantic class Quantitative concept, *level*—a frequent head noun in the collection of biomedical articles—is not. What complicates things further is that endpoints represent the processes, mechanisms, activities that are happening on the molecular, cellular, tissue, organ, or body level and as such can span semantic classes or be comprised of several semantic classes. By their nature, endpoints represent very specific processes and are typically expressed as a compound noun. In the sentence above, *plasma insulin concentration*, *pancreatic insulin content* and *beta-cell mass* were identified as endpoints. Matching *plasma insulin concentration* to its semantic class would fall under the category of a complex match as *plasma insulin* would be matched to one concept and *concentration* to another. What complicates things further is the fact that *concentration* also represents the case of overmatching because it is identified as a Mental Concept but also as a Quantitative Concept. It is the surrounding context that can determine the concept that should be used for *concentration* which in this case is a Quantitative Concept. *Pancreatic insulin* matches to Neoplastic Process which is not an ideal match for *pancreatic insulin content*. Ideally, we would have *pancreatic insulin content* matched to one semantic class that would be identified as the measurement of insulin in the pancreas. It is the head noun *content* in this compound noun that adds this quantitative quality to *pancreatic insulin* and steers the meaning of the noun in the direction of measurement. The situation is somewhat better with *beta-cell mass*, an example of a complex match. *Beta cell* matches to Cell semantic type and *mass* to Quantitative Concept semantic class, both of which identify the parts correctly.

Most often, endpoints are specific phrases and terms that sometimes indicate the outcome measure and sometimes the property of a compared entity that experienced a change. The following sentence is used to demonstrate the level of endpoint specificity:

(40) Nonfasting plasma glucose levels [endpoint_A] and the overall glycemic excursion [endpoint_B] (area under the curve) to a glucose load [endpoint_B] were *significantly*

[relation modifier] *reduced* [relation] (1.6-fold; P < 0.05) in (Pro_3) GIP-treated mice [entity 1] compared with controls [entity 2]. 16046312

The endpoints in this sentence are *nonfasting plasma glucose levels* and *glycemic excursion to a glucose load*. While *plasma glucose* gets matched to the semantic type Laboratory procedure, *nonfasting* is matched to semantic class Finding which in this case is not ideal. Within the context of a comparison sentence and the information it conveys, the modifier, *nonfasting* provides a very important nuance for the meaning of the entire sentence and it should be retrieved as part of the endpoint.

Multi-document summary reports a number of endpoints that are related to metformin drug comparison to other interventions. That study reported the following endpoints that relate to insulin: *proinsulin concentrations, insulin*, *insulin action*, *insulin concentrations, serum insulin concentrations, insulin sensitivity*, *% suppression by insulin*. The following endpoints relate to glucose: *glucose*, *fasting glucose*, *hepatic glucose production during hyperinsulinemia*, *glucose disposal*, *glucose disposal rate, serum fructosamine*, *glycated hemoglobin (HbA1c)*. These endpoints were grouped based on the main substance they were measuring, insulin or glucose, whereas in this study we grouped the endpoints based on the property that they frequently share: measurement characteristic. The question remains what kind of grouping or matching system is better for the particular and specific nature of endpoints and at what modifier and what level we can start to draw the line. These questions require a medical specialist to intervene and assist with the process of endpoint categorization.

None of the concept matching tools that help extract concepts from clinical notes and electronic records, such as MetaMap (Aronson, 2001), cTakes (Savova et al., 2010), ConceptMapper (Tanenblatt et al., 2010), NCBI Annotator tool (Jonquet et al., 2009), were employed because the problem runs deeper than the choice of an NLP concept matching tool: if the concept is not available in the UMLS it will not get matched (Pratt & Yetisgen-Yildiz, 2003). Further, as indicated in the Noun mapping to an ontology section of Chapter 2, since endpoints are very specific by nature, they will likely not have an exact match in the Metathesaurus as this resource, generally, does not define very specific terms. The problems described above fall under the categories of complex, partial matches, gapped partial matches and overmatching. Most typically, endpoints are very specific phrases and it is this level of specificity that prevents them from being matched to an ontology effectively.

When trying to delineate and understand the concept of endpoint and the nature of endpoints in comparative sentences, the concept of mechanism in the sciences (Machamer, 2000) becomes useful. Mechanisms consist of entities and activities: entities are the things that engage in activities and activities are producers of change (Machamer et al., 2000). Entities, within the context of a comparison sentence, represent things that are compared whereas activities represent the change that has occurred between the entities. The endpoint can be seen as part of the activity process, an entity, a dependent more likely than a continuant (Smith & Grenon, 2004), whose role is to communicate the change that has occurred between the entities. Seen from this perspective, compared entities and endpoints would likely belong to different semantic classes. And, indeed, as we have established so far, not all endpoints lend themselves to Quantitative concept or to measurement. Consider sentence 41:

(41) There is evidence to suggest that the somatic mutational pathway [endpoint_A] may *differ* [relation 1] between invasive [entity 1_A] and LMP ovarian tumours [entity 2_A] and invasive tumours [entity 1_B] are *more likely* [relation 2] than LMP [entity 2_B] to exhibit p53 overexpression [endpoint_B]. 11159743

Somatic mutational pathway is identified as one of the endpoints in this sentence. It is not clear that the concept of mechanism can extend to pathways (Röhl, 2012) but even if this is the case, this type of mechanism and the change that is indicated in the above sentence does not involve measurement of any kind, only the statement that the pathway was different. Clearly, in this case, the endpoint does not lend itself to measurement in the same way as the endpoints that comprise the head noun, such as *concentration, level, degree,* or *mass*. The second endpoint, however, *p53 overexpression,* can be measured. This sentence provides an example where one of the endpoints lends itself to measurement and the other does not indicating that the endpoints, even within the context of the same sentence, do not share the same characteristics.

Given that there does not seem to be an ontology that can be directly applied without any modifications, to test the hypothesis that the semantic class of the head noun of candidate noun phrases can help improve precision and recall with which endpoints and compared entities are extracted from the collection, a locally created dictionary of 91 terms was used. 71 unique terms such as *level*, *concentration*, *rate*, *mass*, *proportion*, and *degree* were categorized as an Amount and 20 terms were used to indicate a population group such as, *control*, *arm*, *trial*, *treatment*. Drugs were also identified as a group because drugs, within the context of comparison sentences extracted from biomedical scholarly articles, are frequently used as a population group that is compared to another group of drugs. Drugs were identified using the UMLS Pharmacologic Substance semantic class. A number of candidate nouns that occur in comparison sentence will not be identified with either Amount or Group semantic class and was assigned a Null value. The feature set from the pilot study was enriched with the information on whether or not the head noun of the candidate noun phrase was more likely to be an Amount or Population Group based on whether the head noun was included in the locally created dictionary of 91 words. If this method proves promising, a subset of an ontology or several semantic classes from the UMLS will be used to indicate a Population Group for applying the model on the entire collection.

3.4.1 Entity 1

The pilot study used a binary classification method, Support Vector Machines algorithm, linear kernel on each of the three comparison facet. This case study was specifically interested in determining whether setting the problem as a multi-class classification might yield better results than a binary classification method. What was also of interest was whether non-linear Gaussian kernel might be better suited for the problem of assigning the right role to the compound noun in the sentence. Table 17 indicates that the results do not improve when multi-class classifier was used and no additional features were added to the model. When four classes were predicted and no additional information was added, precision dropped 0.01 point and recall 0.30 points. When we reduced the number of classes to three (focus on Entity 1 and Endpoint only), the precision increased 0.01 point but recall dropped 0.14 points. However, adding information about whether the head noun of the candidate noun phrase is likely to be categorized as Amount or Group improved precision and recall with binary classifier and linear kernel (BC + A & G). Compared to baseline (BC), precision increased 0.05 points and recall 0.02. This represents the best result with longer sentences. The combination of a multi-class classifier and Gaussian kernel plus additional features (MC$_4$ + Amount and Group) also improved the results. Precision increased from 0.39 to 0.53 (0.14 increase) and recall from 0.47 to 0.57 (0.10 increase). However, given that baseline recall was 0.58 (BC, linear) this actually represents a drop in recall of 0.01 point.

Entity 1 (939 noun phrases, 132 sentences)							Entity 1 (939 noun phrases, 132 sentences)					
Linear kernel							Gaussian kernel					
	BC	MC$_4$	MC$_3$	BC + A & G	MC$_4$ + A & G	MC$_3$ + A & G	BC	MC$_4$	MC$_3$	BC + A & G	MC$_4$ + A & G	MC$_3$ + A & G
Precision	0.38 (0.35, 0.41)	0.37 (0.34, 0.40)	0.39 (0.36, 0.42)	0.43 (0.40, 0.46)	0.49 (0.46, 0.52)	0.50 (0.47, 0.53)	0.39 (0.36, 0.42)	0.48 (0.45, 0.51)	0.38 (0.35, 0.41)	0.49 (0.46, 0.52)	0.53 (0.50, 0.56)	0.56 (0.53, 0.59)
Recall	0.58 (0.55, 0.61)	0.28 (0.25, 0.31)	0.44 (0.41, 0.47)	0.60 (0.57, 0.63)	0.47 (0.44, 0.50)	0.52 (0.49, 0.55)	0.47 (0.44, 0.50)	0.33 (0.30, 0.36)	0.41 (0.38, 044)	0.56 (0.53, 0.59)	0.57 (0.54, 0.60)	0.54 (0.51, 0.57)
F$_1$	0.46 (0.43, 0.49)	0.32 (0.29, 0.35)	0.41 (0.38, 0.44)	0.50 (0.47, 0.53)	0.48 (0.45, 0.51)	0.51 (0.48, 0.54)	0.43 (0.40, 0.46)	0.39 (0.36, 0.42)	0.39 (0.36, 0.42)	0.52 (0.49, 0.55)	0.55 (0.52, 0.58)	0.55 (0.52, 0.58)
Accuracy	0.73 (0.70, 0.76)	0.77 (0.74, 0.80)	0.76 (0.73, 0.79)	0.76 (0.73, 0.79)	0.79 (0.76, 0.82)	0.81 (0.78, 0.84)	0.75 (0.72, 0.78)	0.80 (0.77, 0.83)	0.76 (0.73, 0.79)	0.80 (0.77, 0.83)	0.82 (0.80, 0.84)	0.83 (0.81, 0.85)

Table 17: Six Support Vector Machines Entity 1 classifier results on longer sentences (>30 and <=40 words). 95% confidence interval in parenthesis.

With shorter sentences that are not longer than 30 words and that typically have fewer candidate nouns, precision improved from 0.46 using binary linear kernel classifier (BC, linear) to 0.68 using multi-class classifier and Gaussian kernel (0.22 increase) (MC$_4$ + A & G) while the recall dropped 0.01 point from 0.63 to 0.62. This represents the best result.

Entity 1 (385 noun phrases, 66 sentences)							Entity 1 (385 noun phrases, 66 sentences)					
Linear kernel							Gaussian kernel					
	BC	MC$_4$	MC$_3$	BC + A & G	MC$_4$ + A & G	MC$_3$ + A & G	BC	MC$_4$	MC$_3$	BC + A & G	MC$_4$ + A & G	MC$_3$ + A & G
Precision	0.46 (0.35, 041)	0.45 (0.34, 0.40)	0.45 (0.42, 0.48)	0.55 (0.52, 0.58)	0.63 (0.60, 0.66)	0.65 (0.62, 0.68)	0.44 (0.41, 0.47)	0.62 (0.59, 0.65)	0.50 (0.47, 0.53)	0.64 (0.61, 0.67)	0.68 (0.65, 0.71)	0.72 (0.69, 0.75)
Recall	0.63 (0.60, 0.66)	0.32 (0.29, 0.35)	0.46 (0.43, 0.49)	0.60 (0.57, 0.63)	0.50 (0.47, 0.53)	0.51 (0.48, 0.54)	0.53 (0.50, 0.56)	0.35 (0.32, 0.38)	0.45 (0.42, 0.48)	0.57 (0.54, 0.60)	0.62 (0.59, 0.65)	0.57 (0.54, 0.60)
F$_1$	0.53 (0.50, 0.56)	0.37 (0.34, 0.40)	0.46 (0.43, 0.49)	0.58 (0.55, 0.61)	0.56 (0.53, 0.59)	0.57 (0.54, 0.60)	0.48 (0.45, 0.51)	0.45 (0.42, 0.48)	0.48 (0.45, 0.51)	0.6 (0.57, 0.63)	0.65 (0.62, 0.68)	0.64 (0.61, 0.67)
Accuracy	0.75 (0.72, 0.78)	0.77 (0.74, 0.80)	0.76 (0.73, 0.79)	0.80 (0.77, 0.83)	0.83 (0.81, 0.85)	0.83 (0.81, 0.85)	0.74 (0.71, 0.77)	0.81 (0.78, 0.84)	0.78 (0.75, 0.81)	0.83 (0.81, 0.85)	0.85 (0.83, 0.87)	0.86 (0.84, 0.88)

Table 18: Six Support Vector Machines Entity 1 classifier results on shorter sentences (<=30 words). 95% confidence interval in parenthesis.

In conclusion, associating the head noun of a candidate compound noun with categories such as Amount and Group improved the precision of identifying Entity 1. Compared to the baseline method, recall did not increase and typically dropped 0.01 or 0.02 points except with binary classifier, linear kernel when additional features were used (BC + A & G). Generally, multi-class classifier with additional features (regardless of the number of classes predicted) raised precision of the classifier while the recall dropped minimally. The best performance was achieved with binary classification method, linear kernel and additional features used and also with multi-class classifier, Gaussian kernel and additional features (see Tables 17 and 18).

Given that a series of 12 classification tasks was conducted and given that six of them did not include Amount and Group information and six did include Amount and Group information the question of interest was whether these apparent differences in the results can be attributed to chance. To establish whether adding the Amount and Group information boosts the performance by chance, a series of matched t-test on the two contrasted groups (BC, MC4, MC3—Linear and BC, MC4, MC4—Gaussian) versus (BC + A & G, MC4 + A & G, MC3 + A & G—Linear and BC + A & G, MC4 + A & G, MC3, A & G—Gaussian kernel) for each of the reported metrics was conducted.

For Entity 1 prediction, the differences for individual metrics—precision, recall, F_1, and accuracy on longer and shorter sentences—could not be explained by chance only and were statistically significant (p < .05).

This difference is indicated in Figure 18 where circles represent the results of the classifiers (only the results achieved with shorter sentences, 385 nouns, <=30 words are visualized in Figure 18) that had the Amount & Group information available and squares represent the result of classifiers that did not have this information available. A consistent pattern with both linear and Gaussian kernel is noticeable where the classifiers that had the additional information available (empty circles) show a better performance (higher precision but also recall, especially noticeable with Gaussian kernel).

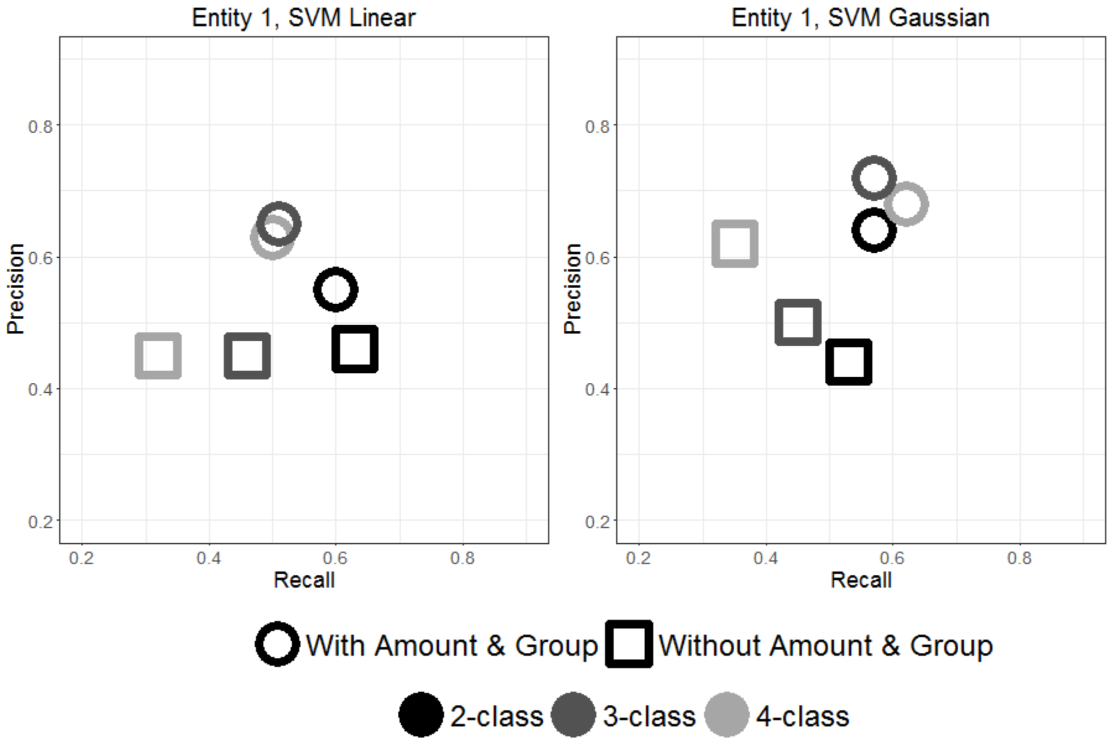

Figure 18: The addition of Amount & Group information improves the Entity 1 metrics significantly (p < .05)

3.4.2 Endpoint

Table 19 indicates the results for endpoint prediction in sentences that are not longer than 40 words. Similarly to Entity 1 classification, setting the problem as a binary or multi-class classifier does not make a difference until information about the type of head noun is added. Such an addition boosts performance with both binary and multi-class classification methods.

More particularly, using multi-class classifier, linear kernel, and additional features (MC$_3$ + A & G) on all 132 sentences improves the precision from 0.42 to 0.56 (0.14 points) and recall from 0.64 to 0.71 (0.07 improvement). Consequently, F$_1$ measure improves to 0.62 (0.09 improvement) and accuracy to 0.79 (0.06 improvement) (Table 19).

Endpoint (939 noun phrases, 132 sentences)							Endpoint (939 noun phrases, 132 sentences)					
Linear kernel							Gaussian kernel					
	BC	MC$_4$	MC$_3$	BC + A & G	MC$_4$ + A & G	MC$_3$ + A & G	BC	MC$_4$	MC$_3$	BC + A & G	MC$_4$ + A & G	MC$_3$ + A & G
Precision	0.42 (0.39, 0.45)	0.42 (0.39, 0.45)	0.37 (0.34, 0.40)	0.45 (0.42, 0.48)	0.47 (0.44, 0.50)	0.56 (0.53, 0.59)	0.39 (0.36, 0.42)	0.42 (0.39, 0.45)	0.40 (0.37, 0.43)	0.50 (0.47, 0.53)	0.47 (0.44, 0.50)	0.45 (0.42, 0.48)
Recall	0.64 (0.61, 0.67)	0.57 (0.54, 0.60)	0.61 (0.58, 0.64)	0.67 (0.64, 0.70)	0.65 (0.62, 0.68)	0.71 (0.68, 0.74)	0.58 (0.55, 0.61)	0.62 (0.59, 0.65)	0.54 (0.51, 0.57)	0.56 (0.53, 0.59)	0.64 (0.61, 0.67)	0.68 (0.65, 0.71)
F$_1$	0.51 (0.48, 0.54)	0.48 (0.45, 0.51)	0.46 (0.43, 0.49)	0.54 (0.51, 0.57)	0.55 (0.52, 0.58)	0.62 (0.59, 0.65)	0.47 (0.44, 0.50)	0.50 (0.47, 0.53)	0.46 (0.43, 0.49)	0.53 (0.50, 0.56)	0.55 (0.52, 0.58)	0.55 (0.52, 0.58)
Accuracy	0.73 (0.70, 0.76)	0.73 (0.70, 0.76)	0.69 (0.66, 0.72)	0.74 (0.71, 0.77)	0.76 (0.73, 0.79)	0.79 (0.76, 0.82)	0.71 (0.68, 0.74)	0.73 (0.70, 0.76)	0.71 (0.68, 0.74)	0.78 (0.75, 0.81)	0.76 (0.73, 0.79)	0.75 (0.72, 0.78)

Table 19: Six Support Vector Machines classifier Endpoint classifier results on longer sentences (>30 and <=40 words). 95% confidence interval in parenthesis.

With shorter sentences (<=30 words), multi-class classifier, Gaussian kernel and additional features improved the results from 0.51 (BC) to 0.58 (MC$_3$ + A & G) and 0.59 (MC$_4$ + A & G). Similarly, recall improved from 0.69 (BC) to 0.74 (MC$_3$ + A & G). Consequently, F$_1$ measure and accuracy increased to 0.63 and 0.78 (MC$_3$ + A & G) and 0.65 and 0.78 (MC$_4$ + A & G) (see Table 20).

	Endpoint 385 noun phrases, 66 sentences)						Endpoint (385 noun phrases, 66 sentences)					
	Linear kernel						Gaussian kernel					
	BC	MC_4	MC_3	BC + A & G	MC_4 + A & G	MC_3 + A & G	BC	MC_4	MC_3	BC + A & G	MC_4 + A & G	MC_3 + A & G
Precision	0.51 (0.48, 0.54)	0.51 (0.48, 0.54)	0.45 (0.42, 0.48)	0.53 (0.50, 0.56)	0.58 (0.55, 0.61)	0.52 (0.49, 0.55)	0.48 (0.45, 0.51)	0.51 (0.48, 0.54)	0.49 (0.46, 0.52)	0.63 (0.60, 0.66)	0.59 (0.56, 0.62)	0.58 (0.55, 0.61)
Recall	0.69 (0.66, 0.72)	0.59 (0.56, 0.62)	0.66 (0.63, 0.69)	0.67 (0.64, 0.70)	0.67 (0.64, 0.70)	0.75 (0.72, 0.78)	0.60 (0.57, 0.63)	0.66 (0.63, 0.69)	0.59 (0.56, 0.62)	0.58 (0.55, 0.61)	0.67 (0.64, 0.70)	0.74 (0.71, 0.77)
F_1	0.59 (0.56, 0.62)	0.55 (0.52, 0.58)	0.54 (0.51, 0.57)	0.59 (0.56, 0.62)	0.62 (0.59, 0.65)	0.61 (0.58, 0.64)	0.54 (0.51, 0.57)	0.58 (0.55, 0.61)	0.54 (0.51, 0.57)	0.60 (0.57, 0.63)	0.63 (0.60, 0.66)	0.65 (0.62, 0.68)
Accuracy	0.73 (0.70, 0.76)	0.73 (0.70, 0.76)	0.69 (0.66, 0.72)	0.75 (0.72, 0.78)	0.78 (0.75, 0.81)	0.74 (0.71, 0.77)	0.71 (0.68, 0.74)	0.74 (0.71, 0.77)	0.72 (0.69, 0.75)	0.79 (0.76, 0.82)	0.78 (0.75, 0.81)	0.78 (0.75, 0.81)

Table 20: Six Support Vector Machines Endpoint classifier results on shorter sentences (<=30 words). 95% confidence interval in parenthesis.

With endpoint prediction, both precision and recall increase when multi-class classifier and additional information are used and we do not see the precision-recall trade-off as with Entity 1. Both types of multi-class classifiers (MC_3 and MC_4) and both kernel methods, linear and Gaussian, benefit from the addition of Amount and Group features. To illustrate, compared with baseline (BC) precision of 0.42, multi-class Support Vector Machines, linear kernel classifier (3 classes) achieved precision of 0.56 (0.14 increase) while recall went from 0.64 to 0.71. With shorter sentences, multi-class (3 classes) and Gaussian kernel achieved precision of 0.58 compared to 0.51 earlier best result (0.07 increase) and recall of 0.74 compared to earlier 0.64 (0.10 increase).

Similar to Entity 1, the differences between results achieved with or without Amount and Group information could not be attributed to chance. The differences were statistically significant for each individual metric (precision, recall, F_1, accuracy) ($p < .05$).

Figure 19 indicates the difference between the results of the classifiers that had the Amount & Group information available (empty circles) and those that did not (empty squares).

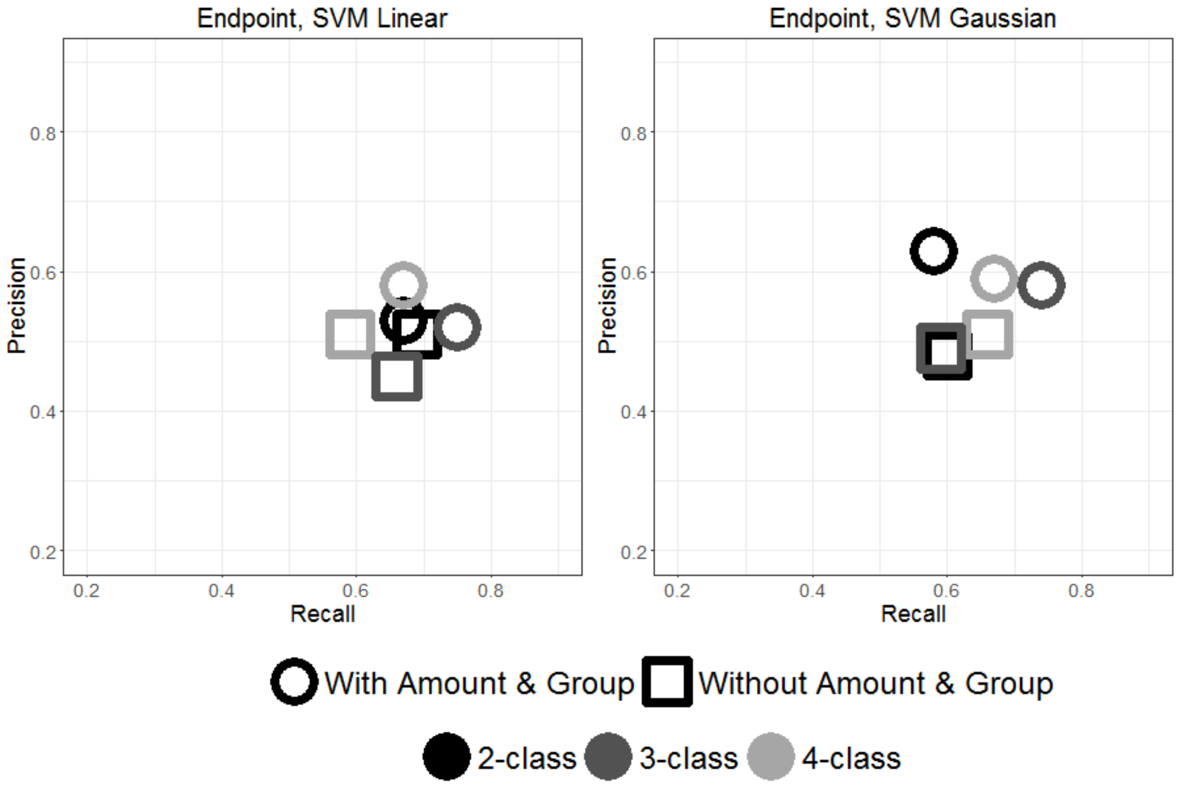

Figure 19: The addition of Amount & Group information improves the Endpoint metrics significantly (p < .05)

Noteworthy in Figure 19 are the results of 3-class classifier with both linear and Gaussian kernel. The 3-class classifier distinguishes between entity 1, endpoint and all the rest of the nouns. Interestingly, the results of this classifier produce the best recall compared to the rest of the classifiers. Future work will dig deeper into this result. Only the results for shorter sentences (<=30 words) are visualized in Figure 19 (see Table 20).

3.4.3 Entity 2

Interestingly, the identification of Entity 2 does not benefit from additional information. As Table 21 indicates, the performance of the linear kernel classifier dropped after the additional information was added with both binary and multi-class classifiers. Earlier work (Blake & Lucic, 2015) reported the closeness of Entity 2 to comparison anchor terms such as *compared with*, *similar to*, and *different from* comprise some of the best indicators for the location of Entity 2. The only improvement in this instance was with respect to recall using binary classifier, Gaussian kernel on all sentences without additional information. In this instance, recall increased to 0.83 from the earlier value of 0.80. This increase, however, was accompanied with a drop in precision: 0.66 compared to 0.74. In conclusion, the addition of the new features boosted the performance for Entity 1 and Endpoint but not for Entity 2 that already boasts a high level of precision and recall. (0.74 precision and 0.80 recall with longer sentences). The implication is that a different set of features is needed to improve the performance of Entity 2 classifier and that shattering of the search space (multi-class classifier) for Entity 2 was not helpful and did not result in better prediction results.

Entity 2 (939 noun phrases, 132 sentences)					Entity 2 (939 noun phrases, 132 sentences)			
Linear kernel					Gaussian kernel			
	BC	MC_4	BC + A &G	MC_4 + A & G	BC	MC_4	BC + A & G	MC_4 + A & G
Precision	0.74 (0.71, 0.77)	0.69 (0.66, 0.72)	0.72 (0.69, 0.75)	0.71 (0.68, 0.74)	0.66 (0.63, 0.69)	0.67 (0.64, 0.70)	0.67 (0.64, 0.70)	0.69 (0.66, 0.72)
Recall	0.80 (0.77, 0.83)	0.70 (0.67, 0.73)	0.79 (0.76, 0.82)	0.69 (0.66, 0.72)	0.83 (0.81, 0.85)	0.78 (0.75, 0.81)	0.80 (0.77, 0.83)	0.74 (0.71, 0.77)
F_1	0.77 (0.74, 0.80)	0.69 (0.66, 0.72)	0.75 (0.72, 0.78)	0.70 (0.67, 0.73)	0.73 (0.70, 0.76)	0.72 (0.69, 0.75)	0.73 (0.70, 0.76)	0.72 (0.69, 0.75)
Accuracy	0.91 (0.89, 0.93)	0.89 (0.87, 0.91)	0.91 (0.89, 0.93)	0.90 (0.88, 0.92)	0.89 (0.87, 0.91)	0.89 (0.87, 0.91)	0.89 (0.87, 0.91)	0.90 (0.88, 0.920

Table 21: Four Support Vector Machines Entity 2 classifier results on longer sentences (<=40 words). 95% confidence interval in parenthesis.

With shorter sentences, recall also increased using Gaussian kernel and binary classifier with no additional features (BC) (0.87 compared to 0.83). However, this was accompanied with a drop in precision from 0.77 to 0.71 (Table 21). In conclusion, Entity 2 prediction does not benefit from additional information and setting up the problem as a multi-class classification did not bring any improvement over binary classifier (BC). A spike in recall was recorded with Gaussian kernel and binary classification method and no additional features.

Entity 2 (385 noun phrases, 66 sentences)					Entity 2 (385 noun phrases, 66 sentences)			
Linear kernel					Gaussian kernel			
	BC	MC$_4$	BC + A & G	MC$_4$ + A & G	BC	MC$_4$	BC + A & G	MC$_4$ + A & G
Precision	0.77 (0.74, 0.80)	0.76 (0.73, 0.79)	0.72 (0.69, 0.75)	0.75 (0.72, 0.78)	0.71 (0.68, 0.74)	0.69 (0.66, 0.72)	0.69 (0.66, 0.72)	0.75 (0.72, 0.78)
Recall	0.83 (0.81, 0.85)	0.78 (0.75, 0.81)	0.82 (0.80, 0.84)	0.78 (0.75, 0.81)	0.87 (0.85, 0.89)	0.83 (0.81, 0.85)	0.82 (0.80, 0.84)	0.82 (0.80, 0.84)
F$_1$	0.80 (0.77, 0.83)	0.77 (0.74, 0.80)	0.77 (0.74, 0.80)	0.76 (0.73, 0.79)	0.78 (0.75, 0.81)	0.75 (0.72, 0.78)	0.75 (0.72, 0.78)	0.78 (0.75, 0.81)
Accuracy	0.92 (0.90, 0.94)	0.91 (0.89, 0.93)	0.90 (0.88, 0.92)	0.90 (0.88, 0.92)	0.90 (0.88, 0.92)	0.89 (0.87, 0.91)	0.90 (0.88, 0.92)	0.91 (0.89, 0.93)

Table 22: Four Support Vector Machines classifier results on shorter sentences (<=30 words). 95% confidence interval in parenthesis.

Figure 20 indicates a larger degree of overlap between different classifiers for Entity 2. Only the results for shorter sentences are represented visually in Figure 20. Adding the Amount & Group information, did not improve the results, if anything, it may have hurt the results slightly.

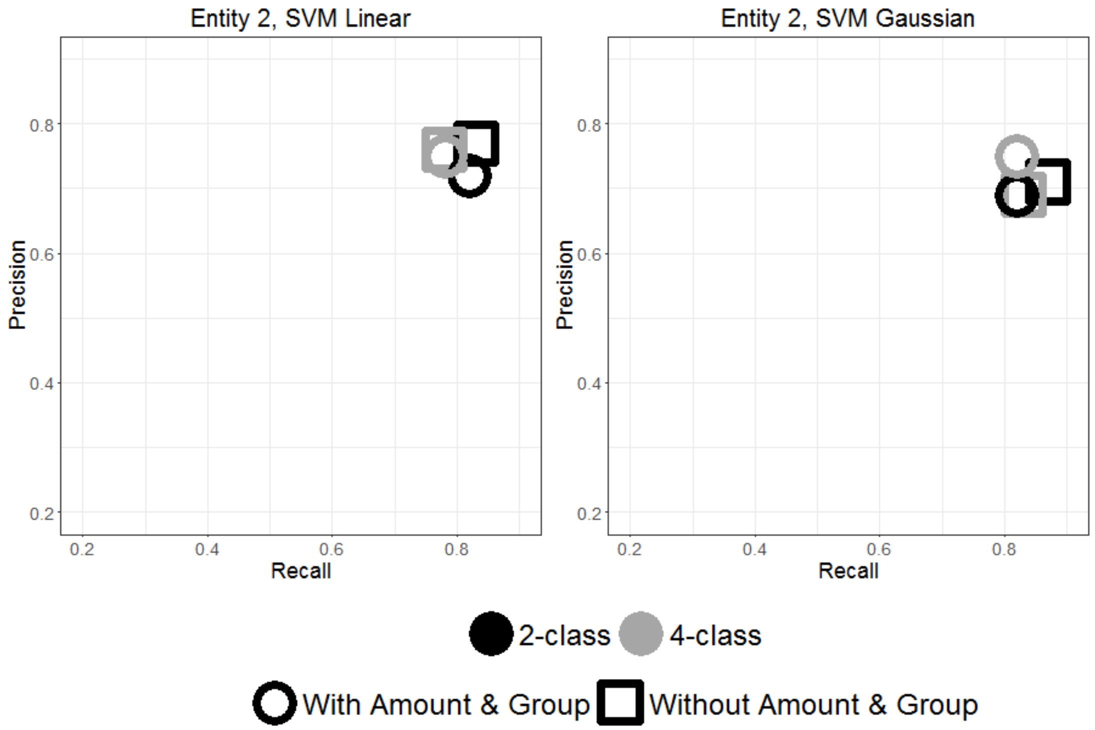

Figure 20: The addition of Amount & Group information does not improve the metrics for Entity 2

It has long been acknowledged that the way in which the scholars write and the way that ontologies are created are not aligned with one another. Word ambiguity, context of the article, precision of grammatical and semantic parsers are standing in the way of better alignment of the free form of textual information in scholarly articles and with entries in ontologies and their definitions. Commonly, a large number of pre-processing tasks is needed in order to convert the text of scholarly articles to a format in which it can be matched to an ontology to enable semantic processing of the text. This study demonstrated that the identification of crucial facets of comparison sentences has benefitted from additional information about the meaning of the candidate noun. Future work will strive to examine the role of the Quantitative Concept UMLS semantic class in assisting with the process of identification, retrieval and definition of endpoints. Also, given that not all endpoints lend themselves to measurement (for example, *pathway*) future work will need to establish other possible ways of modeling endpoints and establishing their significant properties that that can assist in more effective identification and retrieval.

When the results of the experiments that use the additional information were compared to the results that do not use additional information were contrasted it was not clear that these individual differences cannot be attributed to chance only—the difference between these two groups for each individual metric (precision, recall, F_1, accuracy) was *not* statistically significant (P > .05).

What this analysis has demonstrated is that the additional features that indicate whether the head noun of the candidate noun phrase can be categorized as Amount or as Population Group improves the accuracy of endpoint identification and retrieval. These improvements will become especially useful when we try to establish the semantic connection that tie the three facets together and establish the predicate or the nature of comparative relation which is the subject of Chapter 5. The next chapter is dedicated to the discussion of the results of the revised model and its implications. Also, this chapter presents the results of applying the model on the new and previously unseen collection with the aim of understanding to what extent the model built for one purpose and with one collection in the background can be applied to previously unseen collections. Among other things, this will help elucidate the question of how the structure of comparative sentences differ among different collections and depending on the discipline and field. Also, whether the model built for the type of comparative sentences that are found in the biomedical collection of articles can be useful with comparative sentences that come from different genres. In other words, Chapter 4 seeks to answer the question of how discipline specific are comparative structures, how context determines their nature and how some of the characteristics of comparative structures and their lexical, syntactic and semantic features remain stable across disciplines and fields.

CHAPTER 4: MAIN EXPERIMENT

This chapter will summarize the changes that were made to the model with respect to the pilot study and also demonstrate the results that were achieved after these changes were introduced. Finally, the models built are applied on a previously unseen collection of articles—on the topic of breast cancer—with the aim of establishing the extent to which the models can be applied to a new, previously untreated set of articles.

Chapter 3 discussed the heuristics that were considered for improving the model. In addition to focusing only on particular sections of articles and on introducing additional features such as Amount and Group to indicate the semantic group of the candidate noun, the main study eliminated the sentence length restriction that featured in the pilot study. More particularly, the pilot study considered only those sentences that were less and equal to 40 words. In the main study, this restriction was eliminated and the model was applied on all the sentences, regardless of length. Only sentences that had fewer than 10 words were removed because sentences of such short length, typically, do not provide enough space for each of the four comparison facets (two compared entities, endpoint, and relation that binds these three entities) to be communicated. Related to the sentence length restriction elimination, anchor and change term restrictions from the pilot study were also removed (in the pilot study any sentences that had more than 2 anchors or more than 2 change terms were not considered). The set of comparison anchor terms was also revised.

4.1 REVISIONS TO COMPARISON ANCHOR SET

The analysis of 20 Metformin articles from *Diabetes* journal revealed that comparison sentences that were not retrieved using the earlier set of 65 comparison anchors were those that contained the anchors such as *different among*, *similar among*, or *differences among*. This is the reason why these comparison anchors were added to the set of 65 comparison anchors. And yet, an even closer analysis revealed that comparison sentences that contain these anchors, although comparisons in nature, express a slightly different relation than direct comparison sentences considered so far. More particularly, the comparison sentences that contain anchors such as *similar among* and *different among* share certain similarities with superlatives. Consider the following sentence:

(42) There were no differences among treatment groups in the rate of discontinuation or the reasons for discontinuation. 15983221

While this sentence represents a comparison sentence, the relation expressed through this sentence is different from the ones that are the primary focus of this dissertation. Rather than showing a relation *between* compared entities, this relation focuses on a relation that connects, frequently, a group or of a set of entities. In this respect, this type of comparison shares certain similarities with superlatives which compare one entity to a set of other entities, and also expresses the end spectrum of the scale (Sheible, 2007). From computational perspective, superlatives are even harder to process automatically because of the difficulty of establishing the set that an entity is compared to, or, in the case mentioned above, establishing the identity of treatment groups that are being compared. Given that superlatives are not considered in this work, comparison anchors, such as, *different among* and *similar among* were *not* added to the set of 65 anchors. This said, comparison anchors such as *different between* and *similar between* and their variants were kept. Unlike *among* which is a preposition that characterizes a relation between more than two members, *between*—frequently although not exclusively—expresses a relation between two members of a group or between two groups. Consider, for example, the following sentence:

(43) Stimulation [endpoint] of the islets [endpoint] with 3 mmol/l glucose [endpoint] did not show *significant* [relation modifier] *differences* [relation] *between* the wild-type [entity 1] and IRS-1 KO groups [entity 2] (data not shown). 15161756

This sentence uses *differences between* comparison anchor that refers to the two groups compared. The addition of the anchors such as *similar between* and *different between* to the set of comparison anchors is responsible for the retrieval of additional comparison candidate sentences used in the main study. More particularly, there was a 76% overlap between the candidate comparison sentences retrieved in the pilot and main study. The focus on only certain sections of the article in the main study and also the addition of two comparison anchors such as *similar in* and *differences between* are responsible for the 24% difference between the two sets. The precision obtained with a random sample of 1,000 candidate comparison sentences that were obtained using these revised heuristics (the focus is only on select sections) was 76% when direct comparison and anaphoric reference groups were combined. This represents a drop of 10% from the pilot study (86.6% precision). The addition of 5 comparison anchors to the earlier set of 65 comparison anchors widened a pool of candidate comparison sentence and, as Table 23 will indicate approximately 10,000 more sentences were retrieved as candidate comparison sentences. However, as this analysis revealed, there was a trade-off: the increase in retrieval resulted in a drop in precision because not all of the new direct comparison sentences retrieved were relevant. Although *different between* and *similar between* represent comparison anchors that often indicate a direct comparison sentence, sometimes, these anchors are used in sentences that merely express that a comparison was made (sentence 44):

(44) One other major difference between these two strains was identified in this study. 11522683

While comparison anchors such as *similar in* and *different in* are used in sentence that communicate direct comparisons they also help express explicit statements and are not always indicative of direct comparisons.

Table 23 indicates the number of candidate comparison sentences retrieved across three TREC Genomics journals when this revised strategy is employed (no sentence length restriction, focus on the Abstract, Result, Discussion and Conclusion sections, and a revised set of comparison anchors):

Journal	Number of articles	Number of sentences	Candidate comparison sentences	Noun phrases Average number (min, max)	% candidate sentences
Diabetes	2,142	426,743	26,495	8(2, 37)	6.20%
Carcinogenesis	1,958	400,342	10,742	8(2, 51)	2.68%
Endocrinology	5,100	1,212,796	30,374	8(2, 35)	2.50%
Total	9,200	2,039,881	67,611		3.31%
Sentences that are <= 10 words long eliminated			65,980		3.23%

Table 23: New summary statistics for three TREC Genomics journals

What is noticeable is that although the pilot study sentence length and anchor and change term number restriction was eliminated, limiting the number of article sections from which comparative sentences are retrieved resulted in a smaller number of comparison sentences retrieved. The overall percentage of candidate comparison sentences retrieved was 3.31%, compared to around 5.31% in the pilot study. A slight increase in the comparison candidate sentences, however, is noted with *Diabetes* journal where the percentage of candidate comparison sentences increased from 5.41% to 6.23% regardless of the section restriction in the main study. The overall number of candidate comparison sentences in *Carcinogenesis* dropped from earlier 5.01% to 2.68%. In *Endocrinology*, this number dropped from 5.35% to 2.51%. Most likely, an increased number of comparative sentences in *Diabetes* is the result of an increased number of comparison anchors – 70 compared to 65 in the earlier attempt.

4.2 REVISED MODELS APPLIED ON THE ENTIRE COLLECTIONS

In the main study, the training and test sets were combined to create a larger training set consisting of 225 sentences (8 sentences were eliminated). The model built using this training set in the background and the additional features (Amount & Population Group) was applied on all the candidate comparison sentences retrieved from the Abstract, Result, Discussion and Conclusion section of the articles from the aforementioned three journals: *Diabetes*, *Carcinogenesis*, and *Endocrinology*.

Earlier results indicated that the results for entity 1 and endpoint prediction were significantly different based on whether Amount and Population Group information was included. However, given that the earlier study used a dictionary that was created manually and included only 91 terms, it is difficult to establish the true effect and coverage of this feature. To avoid some of the pitfalls associated with a manually created dictionary, and also to take advantage of the UMLS semantic classes, the following two semantic classes were combined: Population Group, Animals (subgroups: Mammal, Fish, Bird, Amphibian, and Reptile), with Drugs as identified through the RxNorm normalized drug naming system. The matching to text was done on the head noun level for two semantic classes (Population Group and Animals). All the unique head nouns from Population Group Mammal, Fish, Bird, Amphibian and Reptile classes were used. If the candidate noun phrase ended with the noun that was in that list, it was assumed that it can be categorized as a Population Group. For Drugs, only drugs that consisted of one (1) word in RxNorm were used. Because RxNorm specifies the name of the drug at different levels of specificity, only term type (TTY) 'IN' and 'BN' were used, as this level of specificity mostly corresponds to the kinds of specificity with which drugs are mentioned in scholarly articles. 'IN' stands for ingredient and implies "A compound or moiety that gives the drug its distinctive clinical properties. Ingredients generally use the United States Adopted Name (USAN)." 'BN' stands for Brand Name and indicates "A proprietary name for a family of products containing a specific active ingredient." Other term types were not included as they typically reference a drug and dose form or a drug and strength.

Another difference from the pilot study is that now the focus is on whether or not a candidate phrase is followed by a preposition or not, rather than on the word that immediately follows and precedes the candidate noun phrase. For this, the Stanford part-of-speech parser (version 3.6.0) was used to indicate whether the candidate noun phrase is followed by a preposition (part-of-speech label 'IN' or 'TO').

The pilot study revealed that sentence length plays a role in the precision with which the model can predict the noun role. The pilot study divided the sentences into two groups: short and long; only those sentences that were shorter than or equal to 40 words were considered. *Short sentences* were those that were less than 30 words long and *long sentences* were those between 30 and 40 words. A key factor that makes it easier to establish the role of the noun in a short sentence is the number of candidate noun phrases in the sentence. Put differently, it makes a difference whether a sentence has 5 or 8 candidate noun phrases: it is more difficult to discriminate among 8 candidate noun phrases than among 4 or 5. To evaluate the effect of the sentence length on model performance more precisely, quartiles for the three journals based on their word length were established. Figure 21 indicates word count quartiles for three journals:

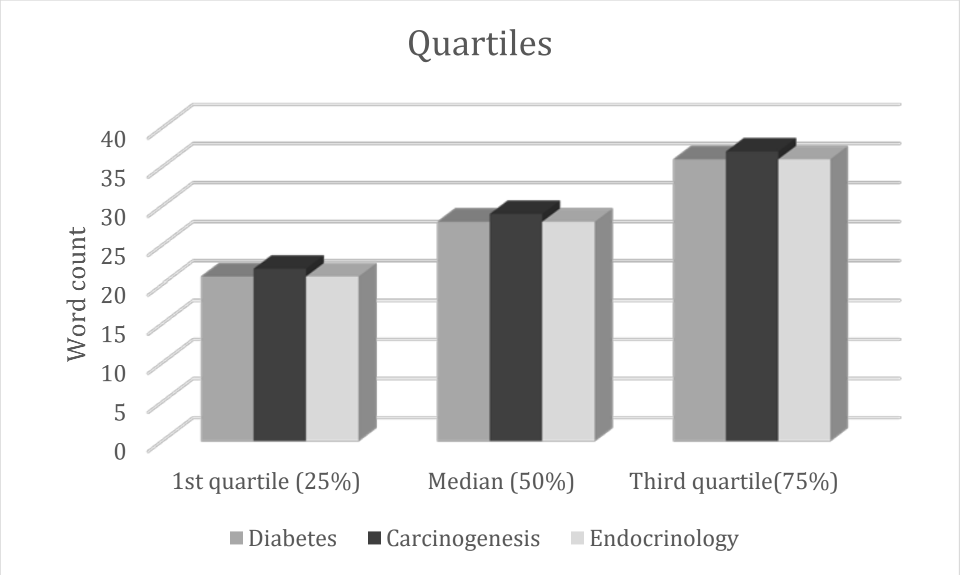

Figure 21: Quartile analysis of comparison sentences based on the number of words

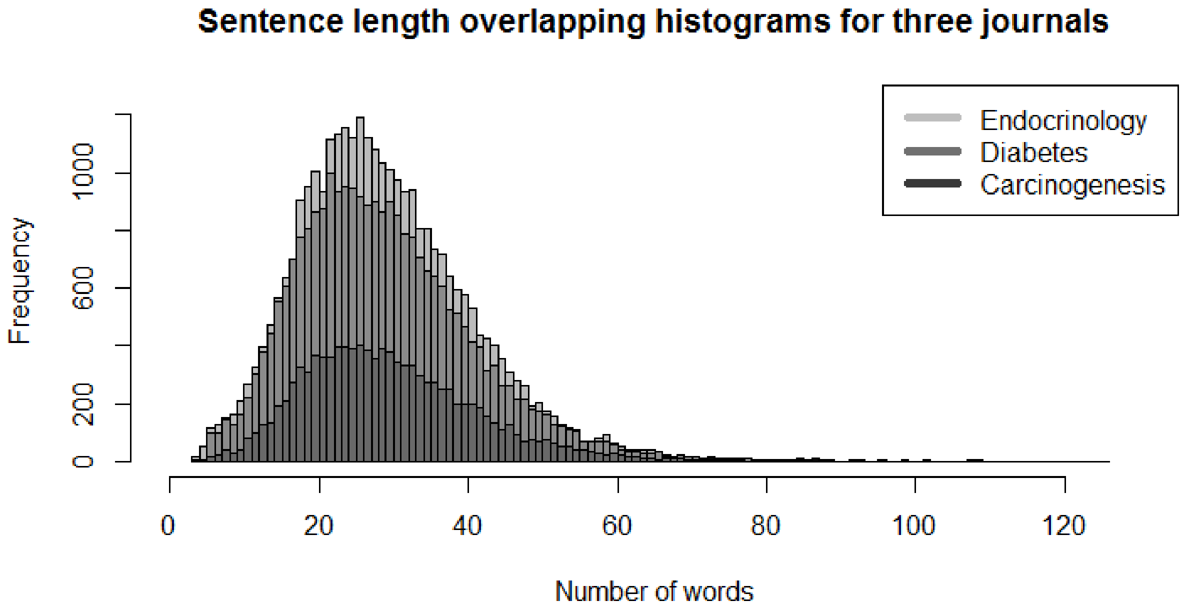

Figure 22: Overlay of sentence length histograms for three journals

Figure 22 overlays sentence length histograms from three journals while Figure 23 indicates a density plot based on the number of words of candidate comparison sentences across three journals. As Figures 22 and 23 indicates, the three journals have similar distribution based on their respective sentence lengths.

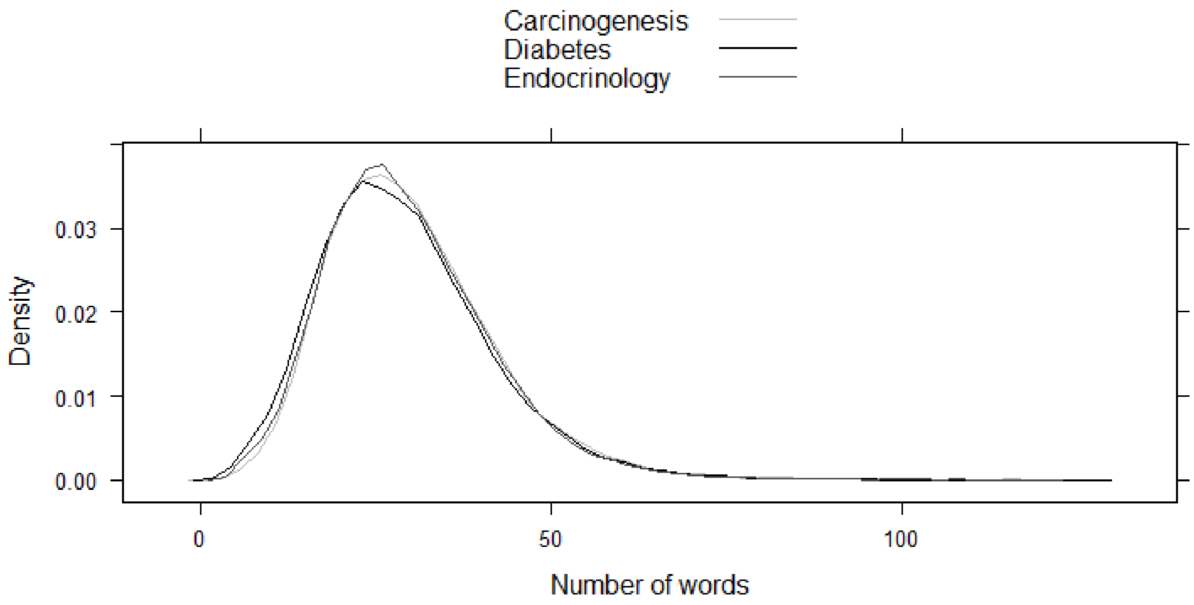

Figure 23: Density plot based on the number of words of candidate comparison sentences

The following graphs show individual word length histograms for each of the journals.

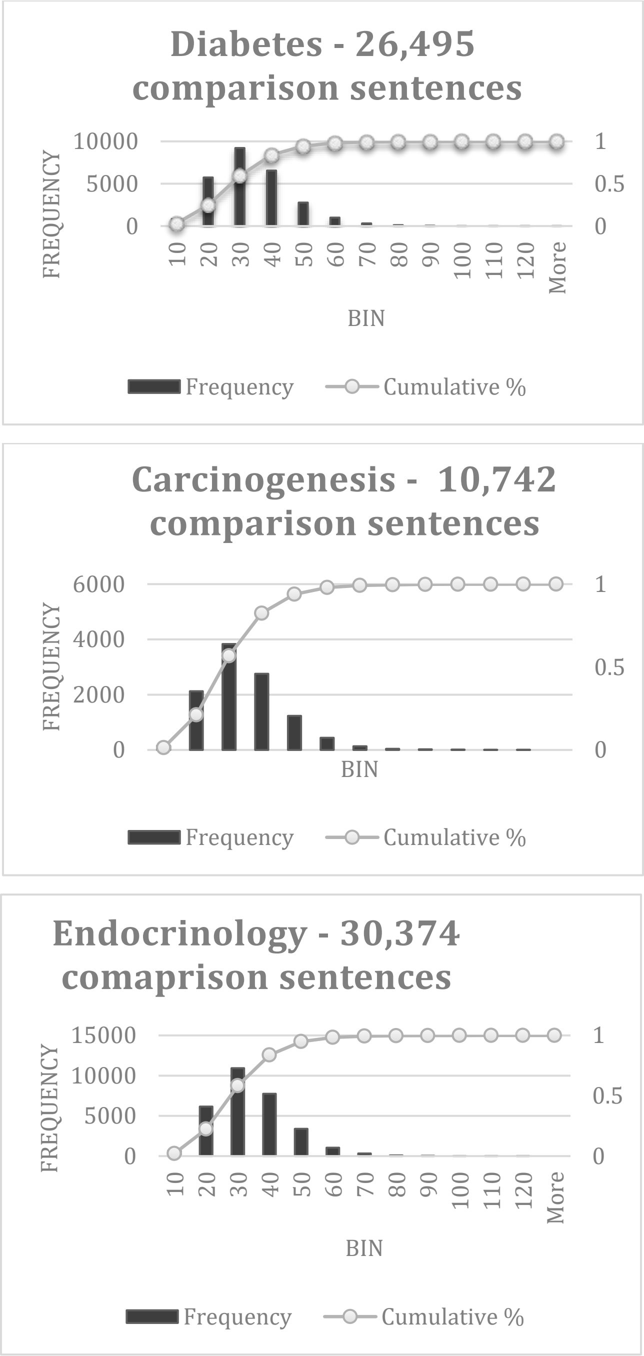

Figure 24: Word frequency histograms

Earlier histogram and density plot analysis (Figures 22 and 23) indicated that the three journals follow a similar distribution pattern based on their respective sentence lengths. Relying on the quartiles as the criterion of sentence length division, the sentences were divided into four groups:

1) short sentences: between 11 and 21 words
2) medium sentences: between 21 and 28 words
3) long sentences: between 28 and 36
4) very long sentences: more than 36 words.

This division allows the examination of the performance of the model against the sentence length and also against the different styles of writing of journals.

4.3 BREAST CANCER COLLECTION OF ARTICLES

At this point, also, a new collection of articles was introduced to evaluate whether the models built can be applied on a collection of previously unseen scholarly articles in the biomedical sciences. A collection of articles dealing with breast cancer was used. A few words need to be said about how this collection was obtained. A collection of full-text articles from the journals listed in Table 24 was retrieved from the PubMed Central database by Joo Ho Lee in October of 2015. The journals were identified by a breast cancer expert and were originally used in a dissertation thesis (Blake, 2003). Originally, 135,554 PMC articles were identified, but only 78,679 had full-text in a suitable electronic format. Subsequently, we narrowed the scope of this journal collection to the articles that had the MeSH subject heading “breast neoplasms.” Only articles that contained this subject heading were included. This narrowed the selection down to 6,552 articles, 748,363 sentences and 76,880 candidate comparison sentences. For comparison purposes, as Table 23 indicates, the number of candidate comparison sentences from three TREC Genomics journals was 67,611 and the number of candidate comparison sentences from the Breast Cancer collection is 76,880, which makes it near-equivalent in terms of size (three TREC Genomics collection had the total of 9,200 articles versus 6,662 in the Breast Cancer collection).

Figure 27 indicates the distribution of sentence length of 76,880 candidate comparison sentences which, clearly shows a similar distribution to candidate comparison sentences in *Diabetes*, *Endocrinology*, and *Carcinogenesis* collections.

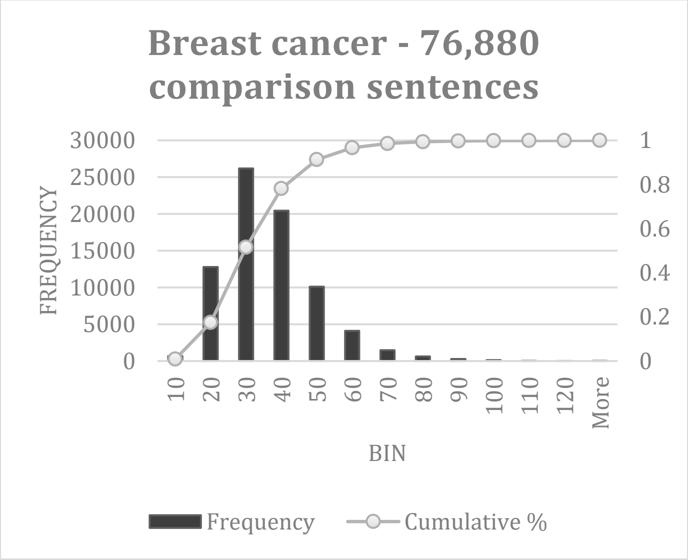

Figure 25: Word frequency histogram for the Breast Cancer collection

Behind the motivation for testing the model on the previously unseen collection of articles was the question of how the model would behave and what kind of results would be obtained. Also, would the model generalize to other collections in the field of biomedical science? A question related to this one is how and whether comparative sentences are determined by the genre or topic or discipline in which they originate in and how much comparative structures across these fields have in common in terms of their lexical and syntactic elements. If the models built are relying, in the large part, on the syntax of a comparative sentence to indicate its main predicates then the discipline differences should not account for much and the models should be able to discern the constituent parts with the same or similar accuracy.

Journal title
American Journal of Epidemiology
The British Medical Journal (BMJ)
Cancer
Cancer Epidemiology
International Journal of Cancer
The Journal of the American Medical Association (JAMA)
The Lancet
The New England Journal of Medicine
Epidemiology
Breast Cancer Research and Treatment
British Journal of Cancer
Cancer Research
Journal of the National Cancer Institute
Journal of Clinical Oncology
Genes Chromosomes
Cancer Lett
Cancer Causes & Control
American Journal of Public Health
Annals of Epidemiology
Carcinogenesis
Human Genetics
BMC Cancer
Breast Cancer Research
International Journal of Breast Cancer
The Lancet Oncology
BMC Cancer

Table 24: Breast cancer journals (Blake, 2003)

4.4 EVALUATION

The revised and improved model was then applied on the remainder of the TREC Genomics collection. To evaluate the new model, 20 sentences from each journal (*Diabetes*, *Carcinogenesis*, *Endocrinology*, and Breast Cancer collection) of four different lengths were randomly selected. As a result, 320 sentences were examined to establish whether they represent direct comparison as defined earlier in the dissertation. Similar to the earlier evaluation, close to 60% of the sentences satisfied this criteria. 10 randomly selected sentences of each length and from each journal were kept which resulted in a sample of 160 randomly selected sentences. The following two models were used and contrasted: Naïve Bayes and Support Vector Machines. Both binary and multi-class classification was used with both models. Additionally, linear and Gaussian kernel of Support Vector Machines classification method were contrasted. Thus, a total of 6 classifier outputs was evaluated. The first step consisted of manually annotating 160 sentences and indicating entity 1, endpoint, entity 2 as well as the relation in each sentence.

To examine the effect of different journals on model performance, the average precision and recall was calculated for each journal which took into account the performance levels for each length and for each of the 6 models. Figure 26 indicates the average recall on the x-axis. The y-axis charts the average precision achieved with four collections where different shapes indicate the particular journal collections.

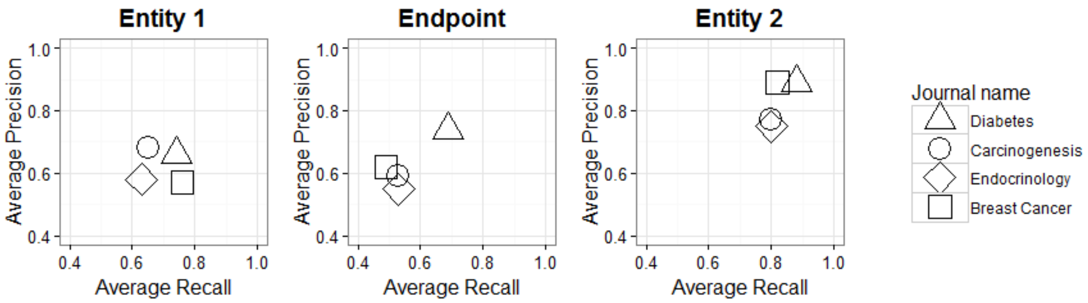

Figure 26: Model performance conditioned on different journal collections

As Figure 26 indicates, the four journals hover close to each other and average precision goes from 0.57 for Breast Cancer up to 0.68 for *Carcinogenesis*. Average recall ranges from 0.63 for *Endocrinology* and goes up to 0.76 for Breast Cancer.

With respect to endpoint identification, Figure 26 indicates that *Diabetes* boasts a generally higher precision and recall than the other three journals which hover in approximately the same area. The *Diabetes* journal achieved 0.74 precision and 0.69 recall. Breast Cancer, *Endocrinology*, and *Carcinogenesis* generally have lower average precision and recall (the average precision ranges from 0.55 to 0.62 and average recall from 0.49 to 0.53).

With respect to entity 2 identification, *Diabetes* and Breast Cancer fare better than *Carcinogenesis* and *Endocrinology*. Both Breast Cancer and *Diabetes* achieved 0.89 average precision versus 0.75 and 0.77 for *Endocrinology* and *Carcinogenesis* respectively. *Diabetes* achieved the average recall of 0.88 versus 0.82 for Breast Cancer and 0.80 for both *Endocrinology* and *Carcinogenesis*. It is worth noting that the new and previously unseen collection of articles—the Breast Cancer collection—achieved almost the same average precision as *Diabetes*.

In general terms, Figure 26 indicates that the models perform similarly well on Breast Cancer, for each of the predicted facets, as they do on the other three journals which were used for model development and testing.

To examine the effect of sentence length on model performance, average precision and recall was calculated against four collections and against each of the six models. Figure 27 indicates average precision and recall conditioned on four different sentence lengths.

Entity 2 shows the clearest pattern where short sentences have the highest average precision and recall which is followed by medium, then long and, finally, very long sentences. The endpoint prediction shows a similar although not as clear a pattern as Entity 2 prediction. With respect to the endpoint prediction, short sentences record highest precision and recall; they are followed by medium sentences, long, and finally very long sentences.

The pattern that was obvious with entity 2 and endpoint prediction is less obvious with entity 1. The graph that shows entity 1 prediction indicates that medium sentences tend to achieve the highest precision and recall and they are closely followed by short, long and very long sentences. Long and very long sentences almost overlap in terms of average precision and recall. Long sentences achieve the average precision of 0.59 and average recall of 0.66, and very long sentences achieve the same average precision and 0.67 average recall.

The assumption of this study is that, in general, identifying the four main facets of comparison sentences is easier with shorter sentences than with longer sentences. The main reason for this premise is that comparison sentences are sometimes buried inside more complex sentences, and disambiguating entity 1, endpoint, and entity 2 roles from a larger set of candidate noun phrases is generally a harder task than disambiguating from a shorter set of candidate phrases. The results indicate a significant overlap between short and medium sentences in terms of average precision and recall achieved and also between long and very long sentences. This may indicate a need for a different division of sentences from the original one: collapsing short and medium sentences into one group and long and very long ones into another.

In general, this analysis revealed that for entity 2 and endpoint identification precision and recall levels generally decrease as sentence length increases and generally increase as sentence length decreases (see Figure 27). This trend was not as visible with respect to entity 1 prediction.

The reason for this lack of a clear pattern that would indicate that sentence length is correlated with the performance that was noticeable with entity 1 may be that this study does not use individual words as the unit of analysis. Rather, it uses noun phrase which often, although not exclusively, comprises several words. While the number of individual words in a very long sentence will necessarily be larger than in a long, medium or short sentence, the number of individual noun phrases in a short, medium, long and very long sentence is usually less varied: a very long sentence may, after all, consists of a very few noun phrases. Thus, a better estimate for model performance may provide the number of noun phrases in the sentence. This analysis is left for future work.

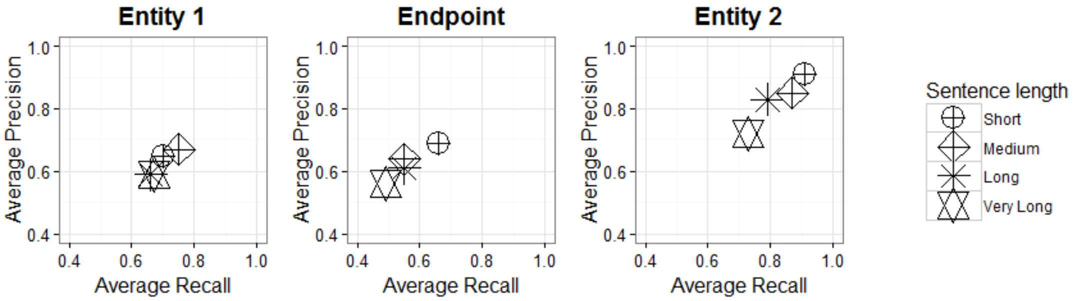

Figure 27: Model performance conditioned on different sentence lengths

Finally, the effect of six different models on the performance was examined by calculating the average precision and recall for the six models regardless of the journal collection in which they originated, and irrespective of the length of the sentence. Figure 28 indicates the results of this analysis.

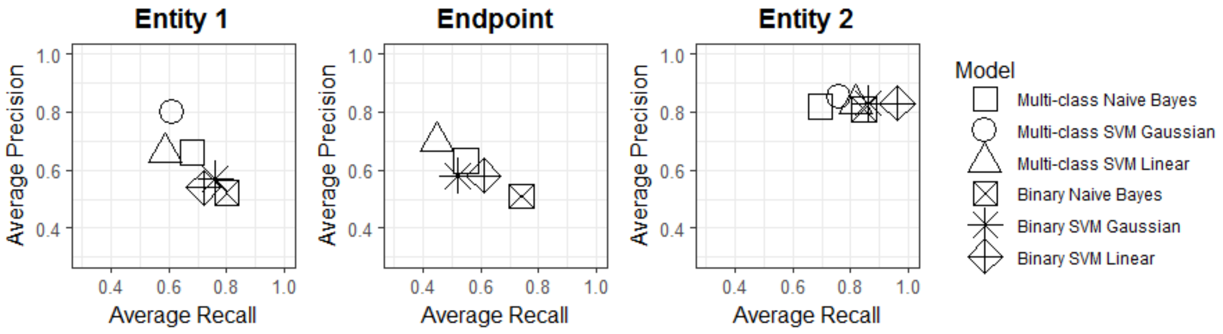

Figure 28: Model performance conditioned on different classification model

As Figure 28 indicates, multi-class classifiers (empty square, circle, and triangle) achieve a better precision but slightly lower recall. With entity 2 prediction, this trend, however, is not very visible and there is more overlap between different models. Binary classifiers show a slightly higher recall with respect to entity 1 and endpoint prediction, and binary Naïve Bayes classification method results in the highest average recall. It is, however, binary SVM linear classifier results that achieve the highest recall with respect to entity 2 prediction.

In general, entity 1 and endpoint are harder to predict automatically than is entity 2. This is most likely because of entity 2's position in the sentence and its closeness to comparison anchors which turn out to be good indicators of the noun that assumes the role of entity 2 in the sentence.

Depending on whether precision or higher recall is more important, we may choose a different model. Multi-class classifiers operate differently than a binary classifier. The former predicts four labels given one set of candidate nouns. Each method has its own advantages but also disadvantages. In general, the multi-class classifiers result in a higher precision but lower recall level.

The individual panels in Figure 29 indicate how well 6 models predict entity 1 conditioned on different journal and length.

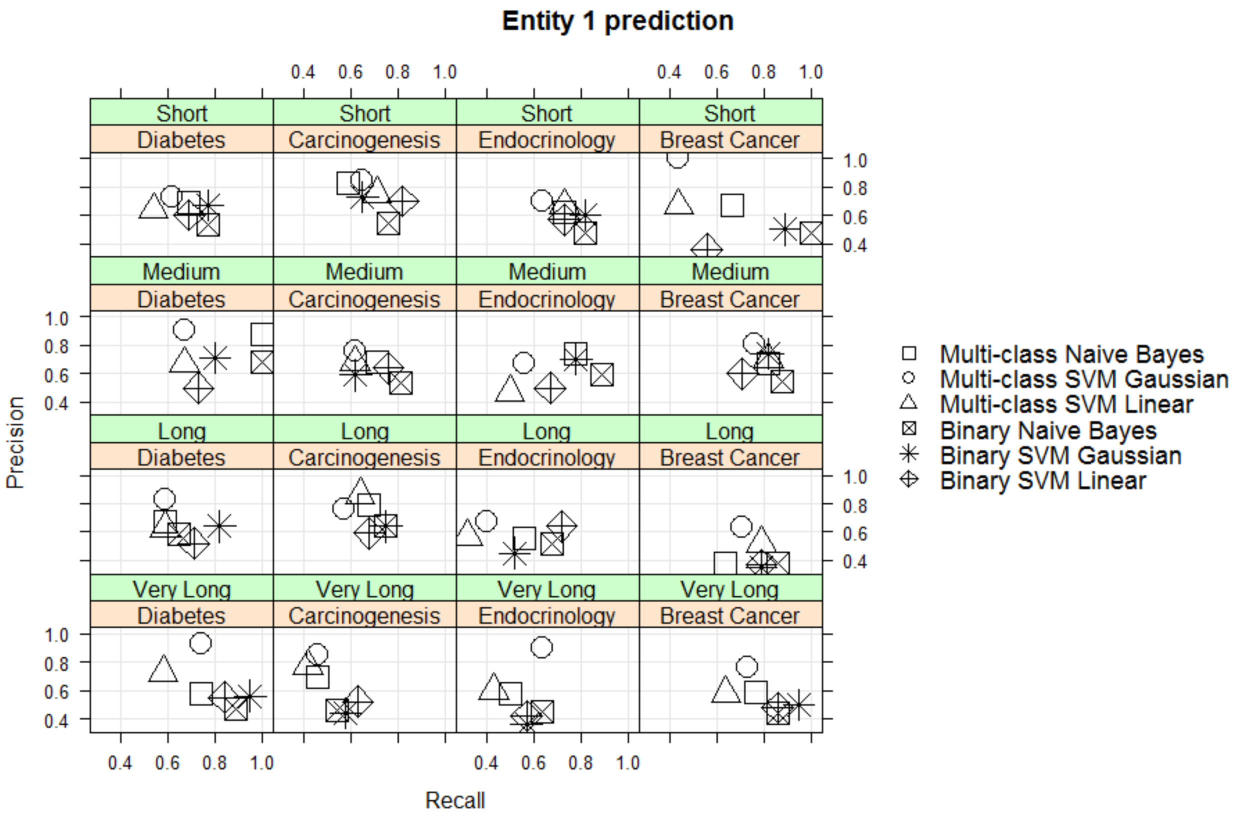

Figure 29: Model performance conditioned on different journals and sentence lengths

As shown in Figure 29, short sentences from four different collections generally show higher precision and recall than do very long sentences, with the exception of the Breast Cancer collection which features a high recall but rather low precision. As the length of the sentence increases—this trend is particularly visible with *Endocrinology* collection of articles—the performance generally decreases and moves to the lower right-hand side part of the quadrant which indicates lower precision and recall levels.

Figure 30 indicates the performance of six different models conditioned on different journals and lengths with respect to endpoint prediction. If we zoom in on the Breast Cancer collection only, it is visible that the performance of each of the models is generally better when applied on short sentences: as the length of the sentence length increases, precision and recall levels generally drop. The *Diabetes* collection reflects a similar trend: the models, in general, perform best when applied on short sentences and the performance generally decreases as the sentence length increases.

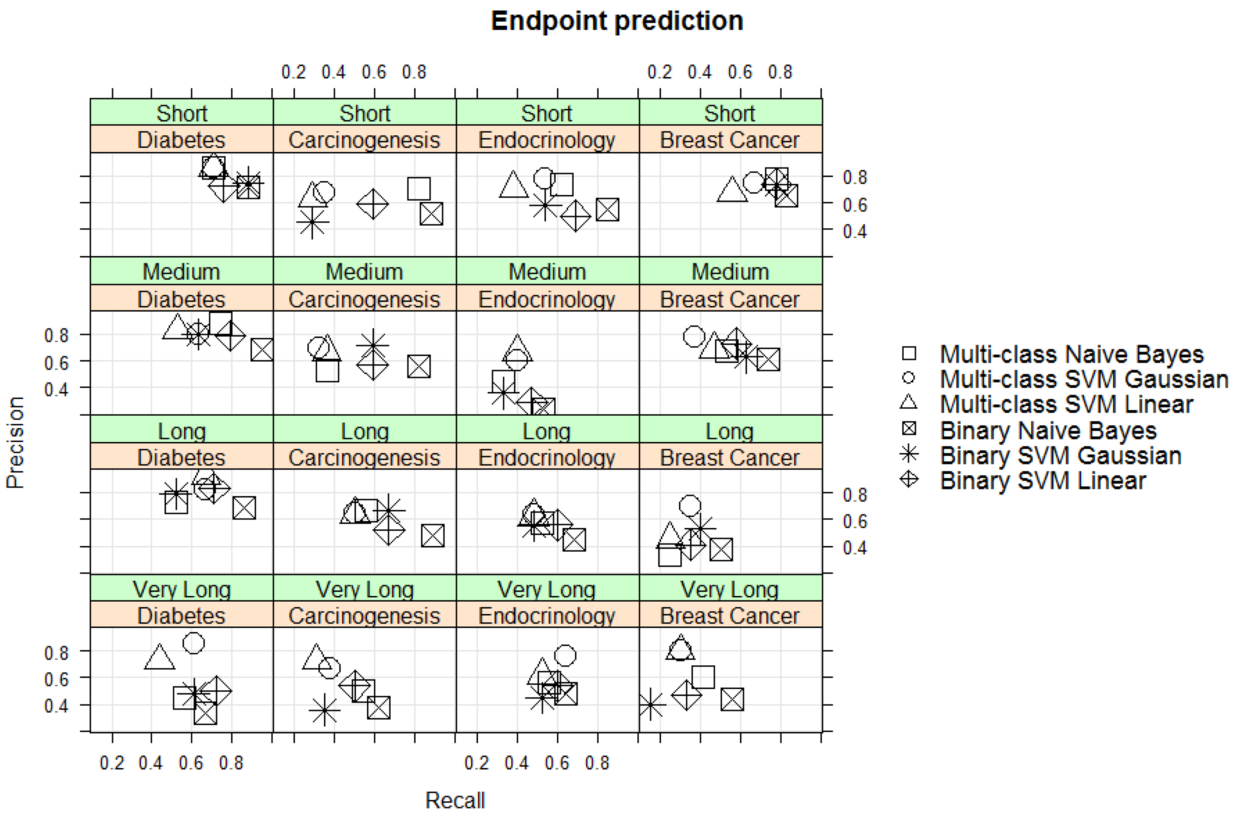

Figure 30: Model performance conditioned on different journal and sentence length

Figure 31 indicates the performance of six different models with respect to entity 2 prediction conditioned on different journals and lengths. The prior assumption that the models will work better when applied on shorter sentences than on longer ones proves valid when it comes to predicting the role of entity 2. While with each of the four collections this trend is apparent it is probably most clearly manifest with the *Carcinogenesis* and *Endocrinology* collections. The panels that show only two or three data points are the result of the overlap of different models.

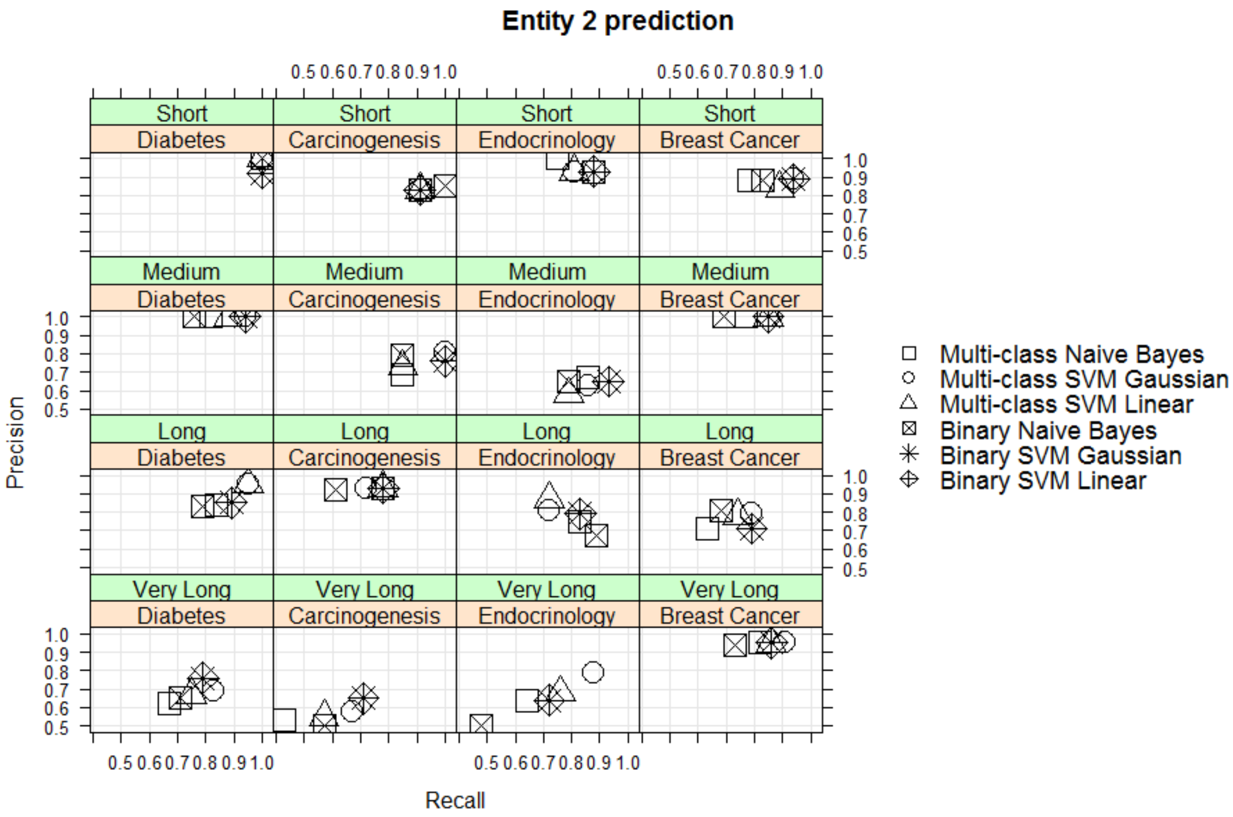

Figure 31: Model performance conditioned on different journal and sentence length

The four plots that demonstrate performance of the models on short sentences show more similar performance than the models applied on long and very long sentences. Not only does the performance generally decrease but it also gets more variable as the length of the sentence moves from short to medium to long and very long. However, in general, there is a greater overlap and less variability in terms of precision and recall achieved with different models with respect to entity 2 than with entity 1 and endpoint prediction. As Figure 31 indicates, the multi-class classifier, in general, tends to result in higher precision and lower recall whereas binary classifiers tend to have higher recall and lower precision levels.

The classifier results with three main facets of comparison sentences indicate that different styles of writing in different journals and different content and registries that they describe do seem to have an influence on the performance of the model. The results also indicate that shorter sentences, in general, achieve higher levels of precision and recall than medium, long, and very long sentences. A closer analysis is needed to establish whether the differences between journals and differences based on different sentence length are significantly different from each other. The analysis of whether these differences are significant and a discussion of how different style and content of the article contributes to these differences is left for future work. Also, establishing differences between different models based on the number of candidate noun phrases rather than the number of word tokens they contain is left for future work.

So far, this study has focused on the identification and retrieval of compared entities and endpoints in comparative sentences. What the results so far have shown is that the sentence length and the type of journal all seem to play a role in the precision and recall rates achieved with respect to predicting correctly the candidate noun role in the sentence. Chapter 5 delves into the identification and retrieval of the fourth facet that is of crucial interest in this work: the main predicate or the main relation of the comparative sentence that indicates how the compared entities and endpoints are related to each other and if and how they changed or remained the same.

CHAPTER 5: RELATIONSHIP IDENTIFICATION

5.1 METHODS TO RECOGNIZE THE NATURE OF COMPARATIVE RELATIONSHIP

The nature of the relationship, within the context of comparative sentences, is tightly connected to what has previously been referred to as the *result* of the comparison. More broadly though, the relationship extraction task can be described as the process of “discovering semantic connections between entities.” This dissertation is only interested in comparison sentences that express the result of comparison. As demonstrated earlier, however, the nature of the relationship that connects the compared entities and the endpoint is more complex than the relation that binds the subject and object in semantic triple. Comparative sentences communicate semantic connections between compared entities and the endpoint or the basis of comparison. As discussed in Chapter 2 of this dissertation, a semantic triple does not represent an adequate data structure for capturing the complex relation that ties the compared entities with the endpoint. An N-quad, a quadruple, or a named graph seems like a more suitable data structure for storing the information extracted from comparative sentences (see section 1.5.1, “Comparison relation”). Furthermore, not all comparison sentences contain the result: sometimes a sentence simply states that a comparison was made between two entities without necessarily expressing the basis on which they were compared or the result of the comparison. Although the sentences that do not identify the nature of change are not of interest in this study the current mechanisms that identify comparisons sentences do not filter such sentences out. It seems as though the ability to infer the nature of the relation between the compared entities and the endpoint in an automatic way would simultaneously provide us with the mechanism that can establish whether or not the comparison sentence contains a result.

To better understand how the result of a comparison that has occurred is expressed in comparison sentences, the comparison sentences that contain a result were compared to the sentences that do not contain a result but merely expresses that a comparison has been made. Consider, for example, sentences 45 and 46:

(45) SERCA2a mRNA levels [entity 1] were compared with nondiabetic levels [entity 1] on a Northern blot. 11916940

(46) ADX + CORT animals [entity 1] had plasma corticosterone concentrations [endpoint] *similar* [relation] to sham animals [entity 2] (6.3 2.8 microg/dl) (P > 0.05). 14633853

Sentence 45 expresses that a comparison has been made between *SERCA2a mRNA* and *nondiabetic levels* but does not provide the result of the comparison. This sentence also does not contain the basis of comparison which can certainly act as a *clue* for our method that the sentence does not contain the result. Sentence 46, on the other hand, communicates that *ADX+CORT animals* were compared to *sham animals* based on the *level of plasma corticosterone concentrations* and that both groups exhibited *similar levels*. In sentence 46, it is the comparison anchor, *similar to*, that is crucial for identifying the nature of the relation between *ADX + CORT animals* and *sham animals*. We may conclude—based on these two examples—that the sentence that contains the mention of two entities that are compared as well as the basis of comparison or the endpoint is more likely to contain the result of a comparison. While this is certainly true in some cases, a sentence can contain each of the three facets and also the comparison marker but still not express the result. Consider sentence 47:

(47) In two recent studies with type 2 diabetic patients, the effects of the PPARgamma agonist troglitazone [entity 1] on insulin signaling [endpoint_A] and action [endpoint_B] were therefore compared with the effects of metformin [entity 2]. 12196460

Despite the presence of three facets and the comparison marker (*compared with*), this sentence does not express the result of the change that has occurred. We may conclude that the presence of a comparison marker—though an important surface level feature for the identification of comparison sentences—does not guarantee that the comparison sentence contains the nature of change or the result of a comparison. And yet, the presence of a comparison marker in combination with the verb that belongs to the class of either change or evidence-based verbs as defined previously seems to be associated with a higher likelihood that the sentence will contain the result. Thus, the combination of a comparison marker in combination with the change expression and/or evidence-based verb might represent one of the heuristics that would help us identify the relationship that binds the compared entities and the endpoint.

5.2 HEURISTICS VERSUS SUPERVISED LEARNING APPROACH

Unlike the task of predicting the entities and the endpoint role in a comparison sentence that used the supervised learning approach, the method proposed for predicting the nature of change in a comparison sentence relies on a set of rules/heuristics.

The assumption of the methods used for noun role identification was that, most often, simple and compound nouns assume the role of compared entities and endpoints. Although, as we saw previously, this is not necessarily always the case, nouns, in general, are the class of words that most commonly takes on the role of a compared entity and the endpoint. If a compared entity or an endpoint consists of more than one compound noun, we would expect the system to identify each of the compound nouns taking on the role of an entity or endpoint. If we proceed with this premise and exclude other types of words, such as verbs, adjectives, adverbs and prepositions (unless they are part of a compound noun) and keep only the noun phrases, we reduce the search space for the candidates that play a specific role in a comparison sentence. When faced with the task of identifying relations in comparison sentences, it seems as though it is the verbs rather than nouns that most commonly constitute the class of words that is likely to indicate a relation. Although verbs typically express relations in a comparison sentence, they are certainly not the only class that is used to communicate the relations that bind the compared entities and the endpoint. Consider the following example where the relation that binds the three entities is expressed through a compound noun rather than a verb:

(48) Only *modest* [relation modifier_A] *reductions* [relation] *(although significant)* [relation modifier_B] were seen in fibrinogen levels [endpoint] in the lifestyle group [entity 1] relative to the metformin [entity 2_A] and placebo group [entity 2_B]. 15855347

In sentence 48 *modest reductions* represents the main aspect of the relation that binds *Fibrinogen levels*, *the lifestyle group* and *metformin* and *placebo group*. Although the verb *seen* also plays a role in expressing the relation that binds these three entities, *modest reductions* provides the key for understanding the change that has occurred between the *lifestyle group*, *metformin* and *placebo group* with respect to *fibrinogen levels*.

Similarly, in the following example, it is the compound noun that holds the key for understanding the relation that binds the three facets:

(49) A *58%* [relation modifier] *decrease* [relation] in mammary tumor incidence [endpoint] was demonstrated in DMBA-treated rats [entity 1] fed 20% less food/day [entity 1] when compared with ad libitum-fed carcinogen treated controls [entity 2]. 9934852

As these examples demonstrate, the assumption that verbs would typically indicate a relation does not hold true in a number of cases. Nature of change, as we saw previously, can sometimes be indicated through a comparison marker, such as in the following example:

(50) B6 LFD mice [entity 1] had body [endpoint_A] and liver weights [endpoint_B] *similar* [relation] to those found in 129 HFD mice [entity 2]. 15855315

Put differently, the comparison marker, *similar to*, holds the key to identifying the nature of the relation. Sometimes, though, it is the combination of a non-gradable comparison marker, such as *similar to*, and a change or evidence-based verb that both express the nature of change that has occurred:

(51) In the present study, we confirmed our previous observation that 2 d [entity 1] of fasting [entity 1] in male rhesus monkeys [entity 1] *triggers* [relation] neuroendocrine responses [endpoint] *similar* [relation] to those observed in healthy men [entity 2]. 11796492

In sentence 51, both *triggers* and *similar to* are responsible for communicating the nature of change that has occurred.

This variety in how the nature of change is communicated in comparison sentences would require a rather large training set to account for the myriad ways of expressing relations. Additionally, we do not restrict the problem to several relations of interest which means that any relation that binds the three entities is appropriate. Given that the creation of a sufficiently large training set that would contain an adequate number of relation instances of various types would require a significant allocation of time and resources, a more practical method for identifying relations consists of developing a set of rules/heuristics that rely on the syntactic structure of comparison sentences in the background.

5.3 DESCRIPTION OF RULES FOR ESTABLISHING DIFFERENT TYPES OF COMPARATIVE RELATIONSHIPS

Three methods for identifying the semantic connection between entities have been identified: knowledge-based, supervised and self-supervised methods (Konstantinova, 2014). Domain-specific tasks are associated with knowledge-based methods and a specific set of relations that are of interest. Systems that use these methods rely on pattern matching rules that are frequently manually crafted. Rich rules, therefore, characterize the identification of relations that connect entities in domain-specific areas using knowledge-based methods. Given that the collection and the relations we are dealing with here are domain-specific and, as of this writing, not identified through an ontology or knowledge base, the approach described here uses a set of rules that were manually crafted. While relationship identification tasks frequently operate within a set of rules for which machine learning is an appropriate method, the question of identification of relations in comparison sentences is an open problem and any relations can be expected. Although certain relations are more common than others (for example, relations such as *increased, decreased*, *stayed the same* or *similar*) reducing the problem to identifying the sentences where one entity *increased, decreased*, or *was similar*—as the results in this chapter will later on reveal—is not sufficient and would eliminate other relations that bind the compared entities in a collection of biomedical articles. This is another reason why a more appropriate method of identifying relations in comparison sentences is the "recognition" of the relation in the sentence based on the presence or absence of certain lexical and syntactic features.

By analyzing the 100 sentences in the training set of sentences from the pilot study, change terms—a list of terms that can denote that a change that has occurred—and syntactic dependencies that commonly surround these change terms as the terms that commonly indicate a relation were identified. These rules were then applied on the set of 132 test sentences.

Comparison sentences are divided into gradable and non-gradable ones; the difference is that the former type of sentences typically involve the use of a gradable adjective. Gradable comparatives, further, order entities according to some aspect whereas non-gradable ones compare aspects but do not grade them. The presence of comparison markers such as *different from, similar to*, *analogous to, identical to* frequently, although not exclusively, indicates that we are dealing with a non-gradable comparison sentence. Change verbs, on the other hand, are most commonly found in gradable comparison sentences. Change verbs help express the gradation between entities and indicate whether a certain property of an entity has increased or decreased. It would seem intuitive that the lack of a change verb and the presence of a non-gradable comparison anchor may indicate a non-gradable comparison. And yet, there are sentences that are characterized by the presence of a non-gradable comparison anchor such as *similar to* or *different from* and a change verb. In these sentences, typically both the comparison anchor and the change verb are responsible for conveying the semantic relation that binds the two compared entities and the endpoint.

The starting point for rule identification of relations that govern comparison sentences is their separation into gradable and non-gradable subtypes based on the type of comparison anchor involved. Diagram represented through Figure 32 represents the graphical form of the rules that govern the separation of comparison sentences into gradable and non-gradable ones and the rules that are used to identify the semantic relation between two compared entities and the endpoint in gradable comparisons.

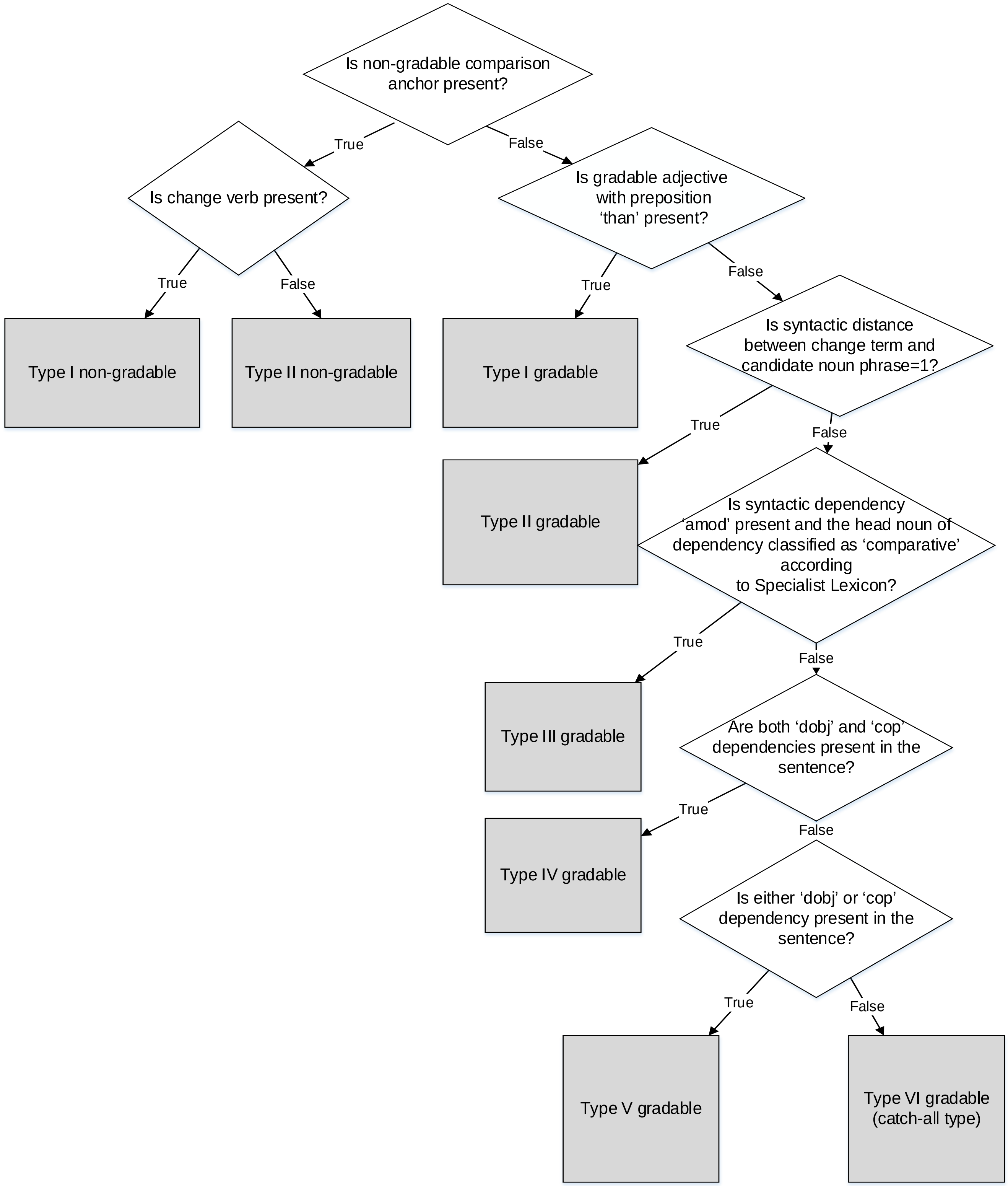
Is non-gradable comparison anchor present?
True
False
Is change verb present?
Is gradable adjective with preposition 'than' present?
True
False
True
False
Type I non-gradable
Type II non-gradable
Type I gradable
Is syntactic distance between change term and candidate noun phrase=1?
True
False
Type II gradable
Is syntactic dependency 'amod' present and the head noun of dependency classified as 'comparative' according to Specialist Lexicon?
True
False
Type III gradable
Are both 'dobj' and 'cop' dependencies present in the sentence?
True
False
Type IV gradable
Is either 'dobj' or 'cop' dependency present in the sentence?
True
False
Type V gradable
Type VI gradable (catch-all type)

Figure 32: Rules diagram for establishing the semantic relationship in a direct comparison

5.3.1 1st rule –identification of non-gradable comparison sentences

If a sentence contains a non-gradable comparison anchor, such as: *different from*, *similar to, analogous to,* and, if it does not contain any change verbs or adjectives, assume that the non-gradable comparison anchor constitutes the semantic connection between two entities and the endpoint. The following examples are illustrative:

(52) Additionally, the maximum inhibition [endpoint_A] in tumor incidence [endpoint_A] and multiplicity [endpoint_B] with Targretin [entity 1] was *similar* [relation] to that achieved with tamoxifen [entity 2]. 10874003

(53) It is interesting to note that in vitro, the affinity [entity 1_AB] and potency [entity 1_AB] of tamoxifen [entity 1_A] and raloxifene [entity 1_B] on the ER [endpoint] is *similar* [relation] to that of 17beta-estradiol [entity 2]. 9389534

In both sentences, it is the comparison anchor *similar to* that indicates the relations that bind the three entities.

To identify non-gradable comparisons, each comparison anchor is identified as being associated with either gradable or non-gradable comparisons. Thus, sentences that contain *different from, similar to* and do not contain a change term would typically be classified as non-gradable comparisons without the change term. When a sentence that contains a non-gradable anchor is identified, assume that the non-gradable anchors hold the key for identifying the nature of the relation.

5.3.2 2nd rule – non-gradable comparison with change terms

If, on the other hand, the sentence contains a comparison anchor that is typically associated with non-gradable comparisons and also contains a change verb, assume that both the change verb and the non-gradable comparison anchor are responsible for communicating the semantic connection between the integral comparison facets. Consider the following sentences:

(54) *Similar* [relation] to the effects of troglitazone [entity 2], metformin treatment [entity 1] *significantly* [relation modifier] *decreased* [relation] HbA_1c [endpoint_A], glucose [endpoint__B], and insulin concentrations [endpoint_C]. 11812753

(55) This work also provided evidence that the mechanisms by which benfluorex [entity 1] *reduces* [relation] hepatic gluconeogenesis [endpoint] are *markedly* [relation modifier]

different [relation] from those of metformin [entity 2], the main antidiabetic compound used in the world. 12145146

The two sentences contain a comparison anchor that is typically found in non-gradable comparative sentences such as *similar to* and *different from* and yet they also contain the verbs that indicate change such as *decrease* and *reduce*. The expression of semantic connection that binds the compared entities to the endpoint is not complete without the addition of the change verb. Sentence 54 expresses that it is not only that *troglitazone* and *metformin treatment* were similar with respect to *HbA_1c, glucose* and *insulin concentrations* but they also reduced each of the mentioned endpoints. Sentence 55 expresses that *benfluorex* is different from *metformin* through the mechanism that reduced *hepatic gluconeogenesis*. Not only are they different but they are different with respect to a specific feature: *the way that they reduce hepatic gluconeogenesis*.

To identify the non-gradable comparisons that contain a change term, a dictionary look up method is used. If a change verb is identified in the sentence, it is assumed that both the non-gradable comparison anchor and the change verb are responsible for communicating the semantic relation that connects the three comparison facets. Needless to say, the accuracy of this method depends on how comprehensive and how complete the list of change verbs used is.

Once the non-gradable comparison sentences and those that use a combination of non-gradable comparison anchor and the change verb have been identified, what is left, the assumption is, is a set of gradable comparison sentences. The ways of identifying semantic relations in gradable comparison sentences is more varied than in non-gradable sentences. The following 4 rules help elucidate the connections that bind the three comparison facets in gradable sentences using their lexical and syntactic features.

5.3.3 3rd rule – gradable adjectives as indicators of semantic relation

If a candidate gradable comparison sentence contains a gradable adjective and preposition *than* that serves the function of a comparison anchor, assume that the gradable adjective is responsible for communicating the association between two compared entities and the endpoint:

(56) HbA_1c [endpoint] was *higher* [relation] in diabetic women [entity 1] than men [entity 2] (P = 0.004). 12031985

This rule is particularly suited for identifying the relations where a single verb is not responsible for playing the role and communicating the semantic connection between two entities and the endpoint. In the above sentence, it is the adjective *higher* that identifies the relation that binds *HbA_1c*, *diabetic women* and *men*.

5.3.4 4th rule – adjectival modifier or adverbial modifiers as indicators of semantic relation

The fourth rule uses the Stanford dependency parser output to help identify the relation that connects the two or more entities and the endpoint. In sentences that contain the adjectival modifier (*amod*) or adverbial modifier (*advmod*) where one of the terms that participates in a relation is indicated as a 'comparative' through Specialist Lexicon, this comparative becomes the candidate for relation identification. Consider sentence 57:

(57) By contrast, GK fetuses [entity 1] *exhibited* [relation_A] *a* higher [relation_A] plasma glucose concentration [endpoint_A] and a *lower* [relation_B] plasma insulin level [endpoint_B] (P < 0.001) as compared with values in Wistar fetuses [entity 2]. 11812746

In this sentence the comparative/gradable adjectives *higher* and *lower* are responsible for communicating the relations that bind *GK fetuses* to *Wistar fetuses* in relation to *plasma glucose concentration* and *plasma insulin level*.

5.3.5 5th rule – close syntactic distance between the change verb and the endpoint as indicator of semantic relation

The fifth rule is characterized by the close syntactic distance between the change terms and the candidate noun phrase. The syntactic distance of 1 between the change verb and the noun phrase signals that the change verb is responsible for communicating the change that the endpoint has experienced. The assumption is that gradable comparison sentences express a change that has occurred and this change is typically communicated with the use of a change verb. The noun that has the shortest syntactic path to the change verb sis assumed to be the endpoint. For example, in sentences 58 and 58:

(58) Mean blood pressure [endpoint] was *reduced* [relation] in response to captopril [entity 1_A] (P < 0.001) or candesartan (P < 0.001) therapy [entity 1_B] as compared to untreated SHRs [entity 2]. 11272159

(59) These results suggested that insulin sensitivity [endpoint] was *increased* [relation] in VPAC2R KO mice [entity 1] compared with their WT siblings [entity 2]. 12239111

It is the change verb such as *reduced* and *increased* that expresses the semantic connection that binds the two entities and the endpoint. However, with the sentences of this type that use passive voice a helpful signal that identifies the semantic connection is the syntactic distance of 1 between the change verb and the endpoint.

Furthermore, in some sentences, the modifier that alters the change verb becomes essential for communicating the connection between the three entities. Consider sentence 60:

(60) Compared with placebo [entity 2], fenofibrate [entity 1] *significantly* [relation modifier] *decreased* [relation] plasma concentrations [endpoint_A] of triglycerides [endpoint_A], total apoB [endpoint_B], apoCIII [endpoint_C], and lathosterol [endpoint_D], as well as the VLDL triglyceride-to-apoB [endpoint_E] and lathosterol-to-cholesterol ratios [endpoint_F]. 12606523

In this sentence, we are not only interested in the fact that *plasma concentrations of triglycerides, total apoB*, *apoCIII*, and *lathosterol* as well as the *VLDL triglyceride-to-apoB* and *lathosterol-to-cholesterol ratios* got *decreased* after *fenofibrate* compared to *placebo* but also that each of them was *significantly* decreased. This is why the modifiers of the change verbs in the sentences of this type become an essential part of the relation that this method aims to extract. However, identifying, in an automatic way, the part of the relation that constitutes a modifier and the part that represents the nature of change or the main part of the relation, is not an easy task. Establishing all constituent parts of a predicate in a comparative sentence (particularly prepositions and modifiers) without which the meaning of the sentence cannot be inferred correctly is left for future work.

5.3.6 6th rule – direct object or copula or both as indicators of semantic relation

Another rule concerns the sentences in which the direct object syntactic dependency acts as the indicator of a relation that connects the two compared entities with the endpoint. Consider sentences (61-63):

(61) The addition of flutamide [entity 1] with testosterone propionate [entity 1] *produced* [relation] a *significant* [relation modifier] *reduction* [relation] in clathrin heavy chain mRNA [endpoint] compared with the effect of supplementation with only testosterone propionate [entity 2] (P < 0.05; n = 5). 9529000

(62) The Av3hGK-treated diabetic mice [entity 1] also *displayed* [relation] a 64.4% [relation modifier] *reduction* [relation] in fasting plasma insulin levels [endpoint] as compared with control groups [entity 2] at week 1 posttreatment [entity 2]. 11574410

(63) Measurement of plasma insulin levels revealed that Av3hGK treatment [entity 1] *resulted* [relation] in a *significant* [relation modifier] *reduction* [relation] in fasting insulin levels [endpoint] in the diabetic mice [entity 1] (66%) compared with both of the control groups [entity 2] (Fig 2B). 11574410

Phrases such as *produced a significant reduction*, or *displayed a 64.4% reduction* in or *resulted in a significant reduction* of become the carriers of the semantic load that connects compared entities and the endpoint. It is not only the change verb or change noun that communicates the essential properties of the relation but the entire phrase communicates the nature of the relation (that is, the meaning) between the entities. Each of these can be identified through the direct object dependency path when other, previously mentioned markers are absent, such as a non-gradable comparison anchor or gradable adjective with the preposition *than*.

The presence of copula in a comparison sentence can also indicate the nature of the semantic connection. Consider sentences (64-65):

(64) After caffeine ingestion [entity 1], plasma FFA [endpoint] was *significantly* [relation modifier] *higher* [relation] compared with placebo [entity 2]. 11872654

(65) There *was* a *significant* [relation modifier] (44%) *increase* [relation] in insulin sensitivity [endpoint] compared with baseline [entity 2] after 26 weeks [entity 1] (P < 0.01) in the rosiglitazone group [entity 1]. 12453903

In these sentence, it is the copula relation that indicates a relation that binds the comparison facets.

Syntactic dependencies *dobj* and *cop* are collapsed into one rule because there exist sentences in which both direct object (*dobj*) and copula (*cop*) relations are responsible for communicating the change that has occurred. The system, therefore, first establishes if the sentence contains both dependencies and if not and there is only a direct object or only a copula relation present it assumes that these dependencies individually are responsible for communicating the relation.

This rule-based system uses a top-down approach that first starts with the separation of a set of comparison sentences into gradable and non-gradable ones. It then proceeds to look for potential candidates that express semantic relations by relying on the presence of certain indicators such as verbs that denote change, comparative adjectives and adverbs and also certain syntactic dependencies (‘direct object’ and ‘copula’). Although these six rules do not exhaust all expression of comparisons, they represent some of the more frequent modes of expressing the relations that bind the three entities found in the collection of the three biomedical journals. For the leftover sentences—the ones that have not been identified by any of the above mentioned rules—it is assumed that the change verb will be responsible for expressing the semantic relation or, if the change verb is not present then the syntactic root of the sentence (which most frequently is the main verb of the sentence).

5.4 RESULTS

The 100 sentences used in the pilot study were annotated and the relation was identified in each. The set of rules described above was then applied on the set of 132 test sentences from the pilot study.

The proposed method seems very successful in recognizing non-gradable comparison sentences or non-gradable comparisons that also use the change verb to communicate the change that has occurred. The total of 21 sentences in the set of 132 (15.9%) were identified correctly and the nature of change was predicted correctly in each of them yielding precision and recall levels of 1.

The rules, however, are not as precise when it comes to recognizing the four gradable types of comparisons. The precision of rule 3, the one that identifies the semantic relation by focusing on the presence of a gradable adjective and the preposition *than* achieves a near perfect precision of 0.97. Only one relation of the total of 31 identified using this method was a false positive relation.

The fourth rule, the one that relies on syntactic distance of 1 between the change verb and the candidate noun achieves a precision of 0.83. Of the total of 36 instances that were identified as containing the syntactic distance of 1 between the change verb and a candidate noun, 30 were truly positive and 6 were false positive.

The application of the fifth rule—which focused on the identification and retrieval of *amod* syntactic dependency and preposition *than* identified 14 instances of which 11 were truly positive.

Of 22 instances for which it was assumed that the direct object relation was responsible for communicating the change that has occurred, 18 were correct and were true positive. Of 16 instances where copula was identified as the communicator of the change that has occurred, 11 were correctly identified.

There were no examples in the set of 132 sentences that contained both direct object and copula relations responsible for communicating semantic change.

The overall precision and recall achieved using this method was 0.86 precision and 0.92 recall. In 11 out of 132 sentences the relation was not identified correctly which accounts for 92% recall. Here is how the frequency of these rules looks like through a pie chart:

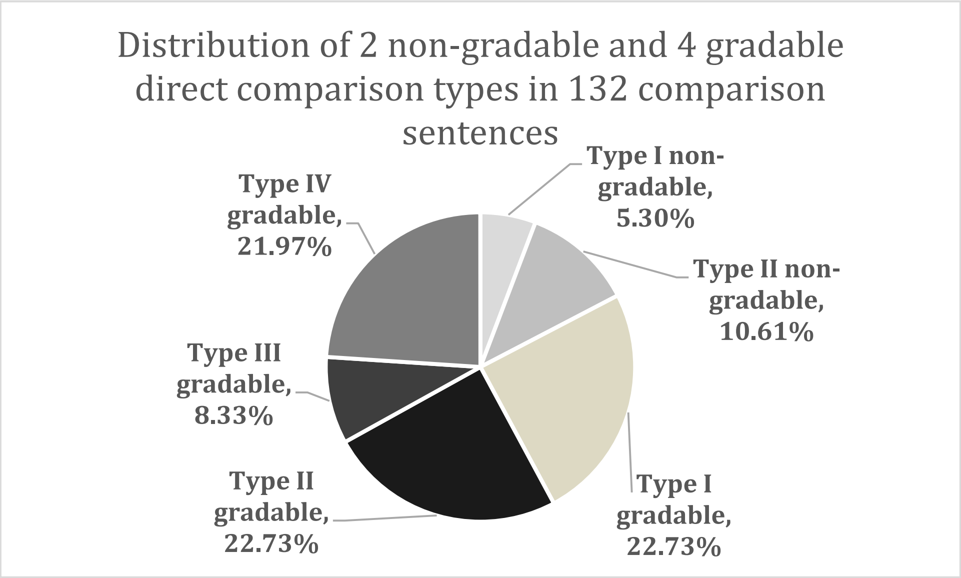

Figure 33: Distribution of 2 non-gradable and 4 gradable direct comparison types

As Figure 33 indicates, in the set of 132 sentences, 17% are non-gradable comparisons. A quarter of the overall number of sentences expresses the semantic relation through the use of a gradable adjective and the preposition *than.* A quarter of the overall number of sentences expresses the semantic relation by relying on syntactic distance of 1 between the change verb and the candidate noun to indicate the change that has occurred. Close to a quarter of the sentences expresses their semantic relation using either direct object dependency, copula dependency or both. 9% of the sentences use the adjectival modifier and the preposition *than* to indicate the semantic relation.

The rules described in the earlier section were applied on the rest of the collection. A more detailed analysis was done with the set of 160 randomly extracted sentences (see section 4.4 “Evaluation”) that were divided into four groups: short, medium, long, and very long. Figure 34 indicates the result of this analysis.

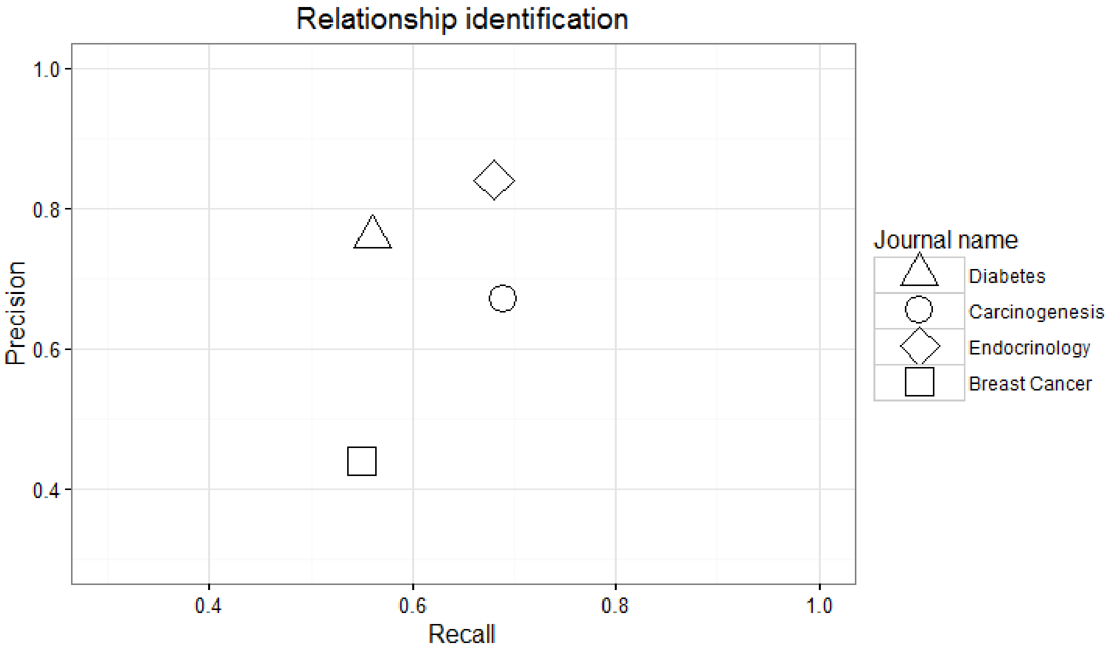

Figure 34: Average precision and recall conditioned on journal

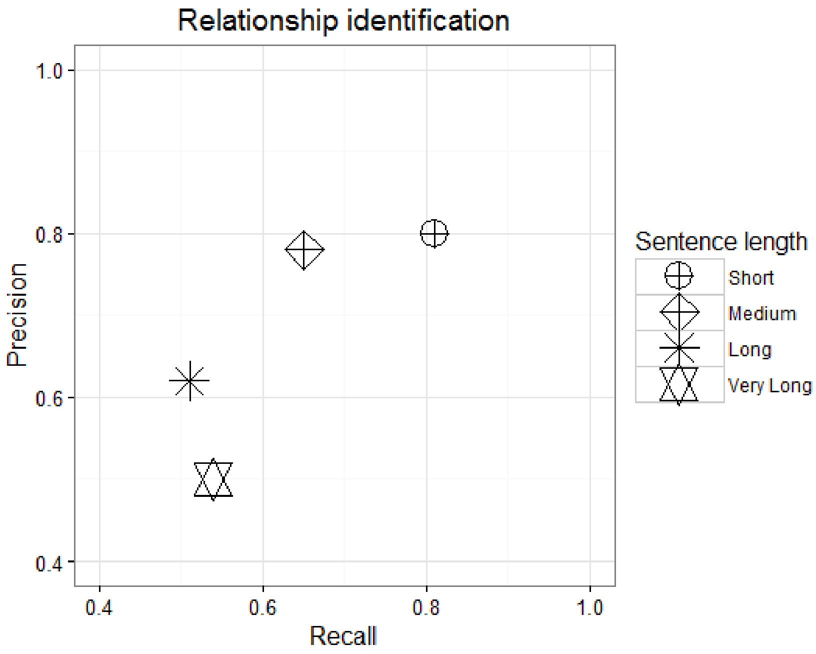

Figure 35: Average precision and recall conditioned on different sentence length

As these two graphs indicate, there are differences based on a) the journal from which the sentences were culled and b) sentence length. In terms of different writing styles found in the three journals, *Diabetes* and *Endocrinology* achieved the highest recall whereas the Breast Cancer collection achieved the lowest recall and precision.

Figure 35 indicates differences based on the length of the sentence with shorter sentences achieving both higher recall and precision (~0.80) and very long sentences hovering in the lower left-hand side of the quadrant and achieving ~50% precision and recall. Figure 35 indicates a clear trend in terms of precision and recall levels decreasing as sentence length increases.

While this finding was not entirely unexpected, it shows a clear trend where the style of the journal (perhaps as a function of the house style guide) and the length of the sentence plays a role in the ability of the rules to identify the relation that binds the three facets under discussion. Short and medium sentences achieve approximately the same precision whereas medium sentences achieve a lower recall. Long and very long sentences which were not considered in the pilot study achieve, in general, lower levels of precision and recall.

5.5 HOW MUCH DO THE RULES HELP (IF AT ALL?)

The rules described earlier were obtained by studying closely a set 100 training sentences from the pilot study. Given that 100 sentences is a limited sample, it is not clear how much these rules might help identify a relation compared to a simpler approach of assuming that the change verb or the nominalized or adjectival form of the change verb will be responsible for communicating the nature of change that has occurred. To establish a baseline and to test whether the rules actually help us identify relation with higher precision, using the WordNet vocabulary, all the synonyms of the change verbs and all their nominalized forms (both singular and plural) were obtained. Also, the lemmatized version of the change verbs has been obtained. The assumption is that nominalized versions of the verbs are sometimes responsible for communicating the change that has occurred. With non-gradable sentences, the approach was similar in that it relied on the nature of comparison anchor to indicate whether the sentence was gradable or non-gradable. Once the rules for non-gradable comparison were eliminated, what was left were four rules that rely on syntactic dependencies in the background to indicate a particular type of a comparison. The question is how these rules help and do they help compared to a simpler approach that assumes that the change term is responsible for communicating nature of change.

To test the validity of developing rules and their ability to indicate a semantic relation, it was assumed that the change term in 120 randomly extracted sentences from three TREC journals (the sample tested here only contains sentences from *Diabetes*, *Endocrinology*, and *Carcinogenesis*) will be the main carrier of the semantic relation that binds the three facets. The achieved precision and recall levels using this approach are contrasted with the rules approach and presented in Table 25:

	Baseline - change terms only	**Syntactic dependency rules**
Precision	0.61	0.76
Recall	0.50	0.64

Table 25: Baseline versus syntactic rules

As Table 25 indicates, the rules bring an improvement of 15% with respect to precision and recall metrics. While the rules are obviously not perfect and do not account for all expressions of relation in comparison sentences, the four additional rules bring an improvement over the assumption that the nature of change will be communicated with a change term. While change terms, on the whole, seem like a good intuitive measure the rules have brought improvement over the baseline.

Most likely, the reason for this improvement lies in the fact that syntactic features get to the structure of comparison sentences and identify a particular type of comparison that uses a particular dependency relation to express a semantic relation. Consider sentences (66-67):

(66) The mean hepatic IGF-I mRNA abundance [endpoint] (IGF-Ia transcript) was *reduced* [relation] by 70% [relation modifier_A] compared with ad libitum-fed controls [entity 1] and by 50% [relation modifier_B] compared with pair-fed controls [entity 2] 9048593

(67) In the PD, hybridization signal intensity [endpoint] is enhanced [relation_A] in the TRH group [entity 1_A] and *significantly* [relation modifier] *reduced* [relation_B] in the T_4 group [entity 1_B] compared to that in controls [entity 2]. 9048604

After studying these two sentences, it becomes obvious that they both use the passive voice to communicate the nature of change that has occurred. While it may appear that the passive voice is an indicator of this type of relation, it soon becomes obvious that this type of relation is not restricted to passive voice. Additionally, passive voice is commonly used in scientific writing, more often than the active voice; over-reliance on this heuristic to recognize this type of relation is not enough. Consider the following sentence that uses the active voice:

(68) Moreover, VOR treatment [entity 1] did not *significantly* [relation modifier] *decrease* [relation] cortical thickness [endpoint] in contrast with orchidectomy [entity 2]. 9165015

The rule that uses direct object or copula dependency to indicate a semantic relation is noteworthy because these rules clearly speak to a pattern of expressing comparison sentences and a particular type of comparison sentences that simple rules would not cover or take into account. Consider, for example, sentence 69:

(69) Furthermore, after IGF-I treatment, insulin-stimulated tyrosine phosphorylation [endpoint_AB] of the IR [endpoint_A] and IRS-1 [endpoint_B] in muscle [endpoint_AB] of the LID mice [entity 1] (Fig 5C and D) was *comparable* [relation] to that seen in the control mice [entity 2]. 11334415

In this sentence, it is copula dependency relation (*was comparable)* that is used to indicate a comparative relation. It is neither the change verb nor the non-gradable anchor that are responsible for communicating the semantic connection but it is the adjective *comparable* used as part of *copula*. By relying on syntactic dependencies to help us identify the nature of change that has occurred we ended up with a sub-type of a gradable comparison sentence: the one that expresses the semantic predicate through the use of copula syntactic dependency. And this is the main advantage of relying on syntactic dependencies to indicate semantic relations in comparative sentences: they allow us to reduce the lexical variety with which comparative predicates are expressed and also identify sub-types of predicates.

The rules for identifying non gradable comparisons rely on the surface level features and on the nature of comparison anchor in the sentence where the presence of a comparison anchor, such as *similar in* or *similar between,* indicates a non-gradable comparison. Four additional rules were developed to identify the relations in gradable sentences *only* and describe 4 sub-types of gradable comparison sentences:

1) The nature of change is indicated through the close syntactic distance between the change verb and the candidate noun phrase.
2) The nature of change is expressed through the use of the adjectival modifier and preposition *than.*
3) The nature of change is expressed using the gradable adjective and preposition *than*.
4) The nature of change is expressed using either direct object dependency (*dobj*) or copula (*cop*) relation or both.

Similar to the comparison facet identification, the relations in the Breast Cancer collection of articles are, in general, identified with lower levels of precision and recall than the relations in the articles from the three journals from the TREC Genomics collections. This difference might indicate a different trend and a different way of describing the nature of change that has occurred with relation to the breast cancer topic vis a vis more varied topics covered in the three journals from the TREC Genomics collection. The results so far suggest that we may need different rules to describe the nature of change that has occurred in the comparison sentences extracted from the Breast Cancer collection of articles than those extracted from the TREC Genomics collection.

5.6 CLAIM IDENTIFICATION

All the analysis conducted so far indicated how well the methods were able to intuit each of the main facets from a comparison sentence, including the main predicate of the comparative sentence. What remained unclear, however, was how well we can identify *each* of the facets that comprises one individual claim. Once again, an individual comparative claim, as envisioned in this dissertation, contains a minimum of two compared entities, the endpoint based on which the comparative entities were compared, and the result of the comparison or the relation/main predicate.

To establish how well the methods developed so far can identify all the relevant facets from an individual claim, each of the individual claims in 160 randomly drawn sentences (described in Chapter 4) was first identified. This analysis identified 228 individual claims. The main facets in the comparative sentences were then identified in each of the individual claims. Very often, one noun phrase corresponds to one of the main facet roles: compared entity, endpoint, or the relation. However, in some cases, two or three noun phrases and sometimes even the entire subordinate clause would take on the role of a compared entity or endpoint. One option is to do a stricter type of evaluation, the one that assumes that all the noun phrases that comprise a role should be identified correctly. If one of the noun phrases that comprises a role is not identified correctly, the prediction is deemed incorrect. Another option is a looser type of evaluation, one in which we are interested primarily in whether or not the classifier at least comes close to predicting the right role for the noun/noun phrases in the sentence. To elaborate, very often, one of the noun phrases takes on the main role within one role and the other noun phrase that is also part of a role, functions as the modifier. If the noun phrase that is not the modifier was predicted correctly and the modifier is not predicted correctly, this would still be deemed a correct prediction. The following type of evaluation represents a looser type of evaluation, and what is of interest here is whether the classifier at least comes close to assigning the right roles to right nouns, even if not every part of the facet is identified correctly.

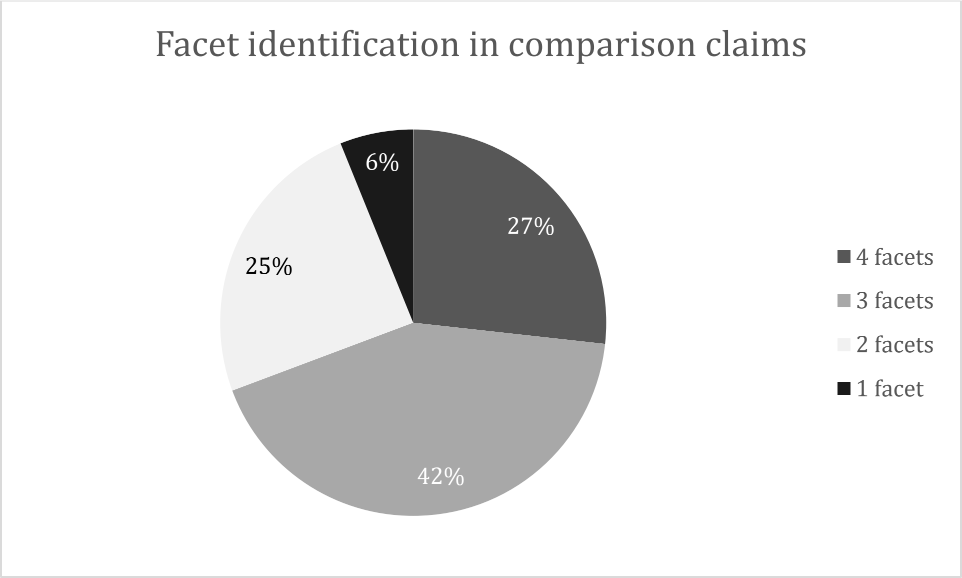

Figure 36: Facet identification in comparison claims

The results indicated that approximately a quarter of all the claims (27%) had each of the four facets identified correctly. The largest number of claims (42%) had 3 facets identified correctly. This means that one of the facets, most often the endpoint or relation, was not correctly identified. A quarter of the claims (25%) had only 2 facets identified correctly. And in 6% of individual claims, only 1 facet was correctly identified (Figure 36).

Of particular interest is the question whether the claims that had each of the four facets identified correctly tend to come from short and medium sentences rather than the long and very long categories. Also of interest was the question of whether sentences from a particular journal/genre of writing might lend themselves to easier prediction of the main facets of comparison than those extracted from the others.

Out of the 61 claims in which all four facets were identified correctly, 30 came from short sentences from three different journals/collections (*Diabetes*, *Carcinogenesis*, and the Breast Cancer collection of articles). Therefore, approximately half of all the claims that had each of the facets identified correctly came from short sentences that are not longer than 21 words. The other half came from the sentences that were of medium, long, and very long length. Obviously, shorter sentences are easier to process and it is easier to identify which nouns play the role of the main comparison facets when a sentence does not contain many candidate noun phrases and multiple individual claims. And yet, as these numbers indicate, the sentences of medium, long, and even very long length were also included in the group that had each of the facets predicted correctly. Although length seems to be a useful indicator, it is obviously not the *only* predictor of success. Future work will establish what other factors play a role in a sentence that has all four facets predicted correctly. For example, it is possible that sentence length and number of noun phrases in the sentence are not perfectly correlated. As mentioned earlier, it is possible for a relatively long sentence to have a relatively small number of noun phrases. Also, establishing the differences between sentences that have all four facets predicted correctly from those where only 1 or 2 facets were predicted correctly is left for future work.

Fourteen sentences had only 1 facet predicted correctly. Out of these fourteen sentences, 6 came from very long sentences (longer than 36 words). 5 sentences come from long sentences (between 28 and 36 words long). Two came from medium long sentences (between 22 and 28 words) and one from a short sentence (< 22 words). Obviously, very long and long sentences seem to be a more likely candidate for having only 1 facet predicted correctly. 97 individual claims out of 228 had 3 facets predicted correctly. These claims come from sentences of various lengths. Approximately a third of the sentences came from short sentences whereas the rest from medium, long, and even very long sentences.

To conclude, identifying all four facets correctly is not an easy task and only approximately a quarter of all individual claims had each of the four facets identified correctly. More complex comparison sentences, such as the ones that use contrastive conjunctions, are harder to process in an automatic way and assign correct roles. One of the conclusions of this work is that it might be useful to implement a pre-processing step that would convert comparison sentences that contain contrastive conjunctions into contrastive *individual* claims before applying the models on them. Developing such pre-processing step, however, is left for future work.

5.7 NEGATION IDENTIFICATION

Thus far, with respect to relationship identification, the focus has been on the main carrier of semantic meaning that connects the three facets. And yet, compared entities, endpoints, and relationships are frequently expressed with the help of modifiers that alter the meaning of the noun phrase. While beyond the scope of this dissertation, the precision with which the models are able to identify the modifier of an entity, endpoint, or a relation is a key factor for understanding how accurately we can piece together information that has been extracted from individual sentences into a comparative summary. The identification of facet and relation modifiers is left for future work although a small entry into this area is made by focusing on how accurately we can establish whether the relation has been negated or not. The difference between one property *increasing* versus *not increasing* or being *different* versus *not being different* is considerable and therefore an important one to be able to capture.

In the sample of 160 sentences, the sentences that contained negation relationship were identified. Out of 160 sentences, 25 were identified as containing the negation of the main relation that is important to be identified alongside the main semantic relation.

Stanford dependency parser output and *neg* dependency was used to identify the expression of negation in the sentence. However, not all *neg* dependencies in a sentence are of interest. Only those *neg* dependencies that connect negation expression and the term that was predicted either as the relationship or as one of the facets (compared entities and endpoints) were utilized. Naturally, the precision with which we are able to identify negation relation will be directly connected to our ability to accurately predict the relationship as well as the three facets.

15 out of 25 sentences had the negation relationship identified correctly and they were counted as true positives. However, there were another 15 instances that were returned as relevant *neg* instances that in fact were not relevant and were therefore counted as false positives. This represents a 0.50% precision. 10 negation instances were not identified and they were counted as false negatives which resulted in 0.60% recall.

Sentences 70-71 represent examples where the rule for establishing negation relation did not work. The most common reason why the negation expression was not recognized was that the sentences expressed negation using less obvious negation terms. Consider sentence 70:

(70) This may in part explain the *lack of* [relation modifier] a *difference* [relation] in initiation [endpoint] we observed between African American [entity 1] and non-Hispanic White women [entity 2]. 16168103

In this sentence, the important modifier for semantic relationship that connects the initiation term and two compared entities (African American and non-Hispanic White women) is the term *lack of a difference*. Ideally, we would want to be able to establish that there *was* a lack of difference in initiation rather than that there *was a difference* in initiation. However, to establish this, Stanford dependency parser and its *neg* dependency were not sufficient as the term *lack* does not trigger the recognition of a *neg* dependency.

The following is another example where Stanford dependency *neg* was not helpful in establishing the true nature of the relationship.

(71) Eight (57.1%) [entity 1_A] of these patients [entity 1_A] had an anthracycline-free interval [endpoint_A] greater_than_ 12 months [endpoint_A], nine (64.2%) [entity 1_B] were classified as having non-anthracycline-resistant disease [endpoint_B], and seven (50%) [entity 1_C] were taxane naïve [endpoint_C], resulting in *no remarkable* [relation modifier] *differences* [relation] when compared with the full patient population [entity 2]. 16168103

Obviously, this is a very long and complex sentence and it is not easy to discern both the relations and compared entities. *Greater than*, *were naïve*, and, *having disease* were returned as candidates for the main semantic relationship that connects compared entities whereas *no remarkable differences*, were not correctly identified. Because the phrase *no remarkable differences* was not identified as the main relation negation expression was not identified correctly.

The following is yet another example in which the fact that the relationship was not identified correctly resulted in negation of the relation not being identified:

(72) In hypoglycemic fed rats [entity 1_A], numbers of Fos-immunoreactive neurons [endpoint_A] in the LHA [endpoint_A] were *slightly but not significantly* [relation modifier_A] *higher* [relation] than in vehicle-treated controls [entity 2_A], whereas the hypoglycemic fasted rats [entity 1_B] had *significantly* [relation modifier_B] *more* [relation_B] Fos_+ nuclei [endpoint_B] than both other groups [entity 2_B], with nearly twice as many as in controls [entity 2_C] (P < 0.001).

This very long and complex sentence, not surprisingly, contains more than one individual claim. Such sentences, in general, represent more difficult cases for predicting comparison facets and relations correctly. Our rules were able to predict only *higher were* as the semantic relation for this sentence rather than *slightly but not significantly more*.

As these examples indicate, the ability to recognize all expressions of negation correctly, including the terms such as *lack of* are important for negation identification. Furthermore, the level of precision with which the main predicate is predicted is directly correlated to the level of precision with which the system will be able to identify negation. Future work will consider other methods for improving the ways in which negation of a comparison sentence can be identified with higher precision and recall.

CHAPTER 6: IMPLICATIONS

6.1 TRACKING COMPARATIVE FACETS IN SCHOLARLY LITERATURE

In addition to answering concrete, comparative information queries, the methods outlined in this dissertation allow us also to connect the scholarly articles across time and analyze how a particular compared entity/endpoint/relation has been discussed over time. For example, we could try to establish which pairs of subject-object (compared entities) have been compared over time and whether there has been a change in the way they have been contrasted. We may also be able to establish contradictions in the studies that discuss the same entities and endpoints but produce different results. Such complex information queries might be of particular interest to biomedical specialists but also to historians, sociologists, and cultural theorists who are interested in tracing the perceptions and attitudes towards a particular nation, historical figure, event, or culture. While current digital library systems certainly allow us to track the mention of a particular concept and also, sometimes, to summarize the information about a particular concept, a more challenging task is bringing together the information that concentrates on not only the two or more entities that have been compared to each other over time but also on the result of that comparison. And herein lies the main contribution of this dissertation: the methods come close to this goal by developing a method to identify the four main semantic elements of comparative sentences (comparative facets in this dissertation).

Not only do the methods as outlined in this dissertation allow us to track the entities that are compared to each other in a collection of articles but they also facilitate tracking of relations or predicates that bind the compared entities. Given that the rules for relation identification, as outlined in Chapter 5, are not restricted by a particular comparison relation but are meant to identify any type of relation that binds two or more compared entities, the method used here has the potential to identify and uncover new relations and also create a typology of relations that occur in comparison sentences. Such a typology of relations would indicate which relations are more common with respect to comparison structures in the collection of biomedical articles and which are rare or indeed very uncommon.

The method outlined in this dissertation lays the foundation for the building of an application. However, constructing an application would imply that this method is vetted and tested by the biomedical specialists in the field that would examine the outputs of the method that can identify not a triple but a quadruple and analyze how the actual utility value of the information obtained. This dissertation envisions a larger study of information behavior of the users of a database such as PubMed. This study would provide users with two outputs: one that current system provides when the user types in a query that involves a comparison relation, and the other that the user obtains when using the method outlined in this thesis. Such a study would indicate the difference in specificity (and thus utility) of information provided to the user.

Finally, another possibility for new research that can develop from the foundation laid out in the present study lies in the area of comparison summary, specifically the tabular format of the summary that would provide the user with the information about the entities that a particular entity has been compared to through columns and the endpoints that the particular entities were compared on through rows. The cells connecting entities and rows would summarize the relations that currently bind the entities and the endpoint. This type of summary would provide a new way of presenting the information and also a new method of scholarly communication through which the authors would inform the readers of larger trends in the collections that they are using or are interested in interfacing with for their research needs.

With respect to the question that was asked earlier in this chapter–on the extent to which current precision and recall levels are able to reveal noteworthy trends and patterns—to answer this question with precision, we would need to conduct a study that would apply the model to a biomedical collection of articles and then sort and aggregate them in quadruples. Such results would then need to be shown to a team of biomedical specialists who could rate the usefulness of the information that the quadruples (two compared entities, the endpoint, and the relation) are able to convey. Future work will aim to produce comparative quadruples in aggregate and will aim to publish them as a way of inviting more discussion on this topic. Although such aggregates of information can be of interest to different groups and even different disciplines, medical or biomedical specialist intervention would be welcome at the moment when the perceived usefulness of the method as outlined in this study and the results it produces are evaluated.

6.2 COMPARISONS IN OTHER DOMAINS

As earlier chapters have indicated, restricting the nature of the entity to be recognized by its semantic class will necessarily bring some false negative results because of the way ontologies are constructed and as a function of genre (the way people tend to construct scholarly articles). These two processes are not aligned and do not have the same goals. This is why not restricting by semantic class the method employed allows the compared entities to, so-to-speak, emerge from the sentence and be revealed by the lexical and syntactic characteristics of the sentence. This, in turn, means that the same method that was applied on the biomedical scholarly collection of articles could, in theory, be adapted and applied on a different collection of articles from a different discipline. And indeed, comparison relations are omnipresent and can be found across different domains. Consider, for example, the following sentence extracted from the journal *Plant Physiology*:

(73) Plants [entity 1] have *substantially* [relation modifier] *higher* [relation] gene duplication rates [endpoint] compared with most other eukaryotes [entity 2]. 2556807

This sentence compares plants with a different species based on their gene duplication rates.

Comparisons are also commonly found in the biological chemistry domain. The following sentence was extracted from the *Journal of Biological Chemistry*:

(74) Mice [entity 1] with a liver-specific knockdown [entity 1] of PERK [entity 1] (produced by AlbCre-mediated deletion of floxed PERK ; see A) showed robust hepatic p-eIF2 [endpoint] to asparaginase [endpoint] that was similar [relation] to AlbCre-negative control mice [entity 2] treated with asparaginase [entity 2] B). 2781691

In the field of nuclear physics—as the following sentence will show—functions and formulas (as opposed to species of plants) can be compared:

(75) The Bethe–Yang function [entity 1] defined in terms of the solutions [entity 1] (106) and (107) to the fermionic Baxter equation [entity 1] (64) Yh(u) = QF −(u) QF +(u), (110) [entity 1] can be brought to the form [endpoint] *similar* [relation] to the one for the hole studied above, Yh(u) = eiPh(u)ShF(u,v) [entity 2]. http://repo.scoap3.org/record/10177/files/main.pdf

Comparisons are also present in the field of scholarly writing about metallurgy. Consider, for example, the following sentence that comes from the article *Diffusion in Silicon* written by Scotten W. Jones:

(76) Within the crystalline grains [entity 1] diffusion [endpoint] has characteristics similar [relation] to bulk single crystal silicon [entity 2]

As these examples indicate, comparison sentences represent a rhetorical structure that is commonly found across many domains. In addition, they are commonly used to report the results of empirical research. As the above examples demonstrate, compared entities will change depending on the discipline, yet the essential structure of direct comparative sentences remains very similar. However, as the results thus far indicated, although syntactic and lexical characteristics of comparative sentences may indeed be helpful for revealing comparative facets across disciplines, the identification of comparative relations in different disciplines—based on the results achieved on the Breast Cancer collection of article—may be more difficult. This particular insight of this work requires more attention from linguists, historians, and different domain specialists. One conclusion of this work is that different disciplines will have their own way of expressing comparative relations for which lexical and syntactic features alone may not be sufficient.

6.3 GOLD STANDARD FOR MAIN SEMANTIC COMPONENTS OF COMPARISON SENTENCES

One of the implications of this study is that the gold standard for identification of compared entities, the endpoint and the semantic relations that binds them, does not exist (see section 1.6). Although the SemRep tool, to a certain degree, could provide a gold standard for this purpose, this tool, as discussed earlier (see section 1.5.1), does not identify the endpoint nor does it recognize all the comparison relations that the methods in this dissertation can engage. Additionally, SemRep predefines entities in advance and restricts them by semantic type, a limitation that method outlined in this study is able to overcome. Even if we were interested in relying on the SemRep output and using it as a form of distant learning of comparison relations the methods and approaches are not sufficiently similar or comparable to allow us to do so. This work, therefore, calls for the creation of a gold standard that would identify the compared entities, endpoints and semantic relation in comparative sentences.

6.4 DIRECT COMPARISON AS A META-RELATION

Probably one of the main implications of this research is that it has focused on one particular form of comparative relation that *cannot* be represented by a semantic triple and that needs at least the fourth contextual element to summarize the relation between the minimum of two compared entities, the endpoint, and the result of comparison. This comparative relation, as Chapter 5 indicated, contains a number of different comparison sub-types: four gradable and two non-gradable ones. Direct comparison thus, the main focus of this work so far, is only one of many comparison types and given the sub-types it contains, it represents a meta-type of a relation in the sense that it represents a higher or more abstract type of relation than the ones that it subsumes—concrete realizations of direct comparisons. The particular structure of a direct comparison sentence can potentially subsume many relations underneath that can be found across domains.

CHAPTER 7: LIMITATIONS AND FUTURE WORK

This dissertation aims to produce the models that can predict the noun phrases in the sentence that take on the role of a compared entity, the endpoint based on which the compared entities were compared against as well as the result or the main semantic relation that connects the entities and the endpoint. As earlier sections demonstrated, second entity boasts largest precision and recall. Entity 1 and endpoint are still sometimes harder to distinguish although improvement has been made since the pilot study by introducing additional features that help identify the meaning of the nouns. A set of rules has been applied that can arrive at the nature of change that has occurred and identify the result of the comparison. In addition to the fact that each of the models built so far would benefit from improvement this study has several limitations that future work will need to address.

The first limitation is with respect to the noun-centric view of comparison sentences: the main comparison facets (compared entities and endpoints) are exclusively represented through single and compound nouns. Although, most often, single and compound nouns are indeed the main comparative facets in a comparative sentence, the method described in this thesis does not take into account prepositional attachment, or, more particularly, prepositions that connect several nouns that often times comprise one comparison facet. Future work will need to take into consideration prepositional attachment to obtain a better, more detailed, and more precise facet representation.

In addition to prepositional attachment, comparison facets, like semantic relationships, frequently take on modifiers which change the meaning of the main noun or the verb that communicates the nature of the compared entity, endpoint or relation. As the earlier sections of this dissertation demonstrated, the variety with which entities and endpoints and relations are described in scholarly literature are the main detriments to a more efficient ontology matching of the scholarly articles. Establishing the boundaries for all the modifiers for each of the comparison facets and the relation is the goal which currently stays out of scope of this thesis.

The second limitation is that this dissertation focuses on only gradable and non-gradable direct comparisons, which on average, appear 63% of the time in a sample of 1,000 randomly extracted comparison sentences. Other types of comparisons that are not considered include superlative relations (see Chapter 1) but also sentences that contain conjunctions such as *while, whereas, although, but,* and also other types of comparisons as described in section 3.2 that are not direct comparisons. This type of sentence would benefit from a pre-processing task that would first establish the contrastive or conjunctive statements and then apply the models described in this study on the individual statements.

In the current system the article sections helped to ensure that the comparative sentences feature a result, but this approach could not be used for free-form text. Although this is currently a limitation of this study, the relationship extraction method described in Chapter 5 of this dissertation looks promising in terms of its potential to narrow down on the sample of interest in this study. Comparison facet identification task in this dissertation has started with the identification of compared entities and endpoints. However, a better way to start this process may be to first identify and extract comparative relations and eliminate all the sentences that do not contain a relation/comparison predicate. This approach might help us narrow down on the sentences that report the result of comparison. Although, this observation is currently included as a limitation of this study, given that relationship extraction results look very promising, future work will first focus on the relation extraction and then expand onto comparison facet identification and extraction.

CHAPTER 8: CONCLUSION

Though comparisons have long been identified as an information need, currently the system that would allow a user to extract comparative information from scholarly literature does not exist. The methods delineated in this dissertation ultimately show that by focusing our attention on individual facets of comparison sentences we can obtain valuable insights into ways in which one entity has been compared over time, into the frequent pairs of compared entities, and into the areas where comparative work has or has not been done. This dissertation highlighted the need for a comparative summary that would focus on the compared entities, the properties of the entities they were compared against as well as the result of the comparison. Although this dissertation does not provide a multi-document comparative summary of biomedical scholarly articles, the methods outlined in this document represent a step in this direction—the hypothesis that has guided this work thus far is that the extracted facets from comparative sentences can be used as content elements in a comparative summary.

One particularly challenging aspect of the facet role identification in a comparative sentence has been the distinction between the first compared entity and the basis of comparison or the endpoint due to the similar context in which they appear. Improvement in this area has been made by adding features that indicate whether the candidate noun represents an amount or measurement or a population group. To illustrate, on shorter sentences, those that are equal to or fewer than 30 words, the F_1 measure increased from 0.46 to 0.58 or 0.60, depending on whether the linear or Gaussian kernel was used with the Support Vector Machines classifier. Multi-class Support Vector Machines classifier using the Gaussian kernel achieved an F_1 of 0.65. With respect to the endpoint prediction, the F_1 measure increased from 0.51 (linear kernel-SVM binary classification) to 0.59 (Gaussian kernel-SVM binary classification) when additional features were used, and to 0.65 when the SVM multi-class classifier was used with the Gaussian kernel. All differences were statistically significant ($p < .05$). For the second compared entity, the additional features did not bring any significant improvement: the performance (measured through the F_1 statistic) dropped marginally from 0.80 to 0.77 when the additional features were added but the difference was not statistically significant ($p > .05$).

Comparison sentences are usually divided into gradable and non-gradable and yet, as this dissertation highlighted, comparison sentences are very complex and as such they lend themselves to different types of categorization. This dissertation has categorized direct comparative sentences according to their meaning and earlier chapters indicated how they can be categorized based on the amount of information or based on the amount of comparative facets they convey. This work has focused specifically on one particular type a comparison sentence: the one that communicates the information about which entities were compared to each other, the basis on which they were compared, and the result of the comparison. It was established that of all retrieved comparisons, only 63% of 1,000 sentences randomly sampled satisfy those constraints. This dissertation, thus, focused on one sub-type of a comparison sentence that is relatively frequent and occurs across domains but has not received enough attention thus far.

When viewed from the point of view of the semantic predicate--the semantic connection that ties the compared entities and the aspect on which they were compared—the analyses conducted in this dissertation identified four sub-types of a direct comparative gradable relation and two sub-types of a direct comparative non-gradable relation. One implication of this finding for future work is that targeting a particular sub-type of a direct comparison may bring better results than targeting all six direct comparison sub-types that this dissertation identified. Future work will thus treat each of these six sub-types of a direct comparison relation as separate relation and will aim to increase the overall precision and recall by developing methods for individual sub-types. This approach may not only help provide better results with respect to comparison facet identification but it can also lead to a more precise candidate direct comparison sentences identification and extraction.

The results indicated that when tested on a random sample of 160 sentences from four difference collections and of four different lengths, the forty sentences of short length achieved the average F_1 measure of 0.80. Those that were termed medium achieved 0.71, long sentences 0.56, and the very long ones, 0.51. A method that seeks to identify the semantic relation has the potential to identify and filter out sentences that do not contain a result – an additional reason why the identification of the semantic connection is of cardinal importance and the comparative facet extraction method should start with the semantic relation extraction and then expand onto other comparison facets (compared entities and the endpoint).

A new collection was created to ensure that the methods created would generalize. The results show that the facet prediction (two compared entities and the endpoint) was comparable to the precision and recall levels achieved with the TREC Genomics collection of articles (*Diabetes*, *Endocrinology*, and *Carcinogenesis* journals). With semantic relation identification, a lower level of precision and recall was observed which might be an indicator of different relations in this collection and different ways of thinking about the results in this collection, compared to the TREC Genomics collection. Future work will examine these differences and try to establish variables that influence how semantic relation is communicated across different biomedical collections and also across other disciplines.

Comparison facet identification in candidate comparison sentences is what guides this dissertation. Comparison sentences sometimes communicate only one claim; frequently enough, though, they express plural claims. Although the end goal of this dissertation is to identify and extract relevant facets from comparative sentences, ideally, the goal is to associate the right facet with the right claim, if the sentence expresses multiple claims. As the individual claim analysis in Chapter 5 indicated, it is possible, at present, to correctly predict each of the facets for a quarter of all the claims in a randomly extracted sample of sentences. In a sample of 160 comparison sentences of different lengths, 228 individual claims were extracted and in 27% of these individual claims the models built were able to discern and identify the four roles correctly. In approximately 42% of all individual claims this information was identified correctly for at least 3 out of four facets. This means that for approximately 70% of individual claims, the models built are able to predict and identify at least 3 out of 4 facets. Although improvement has been made with endpoint identification with the introduction of additional features, this facet still remains the most elusive one.

Sentence length and the journal all exert an influence and have a certain effect on predicting the facets of a comparison sentence. Generally, the method described in this study works better on shorter sentences than on longer ones. Longer sentences as well as complex sentences that contain conjunctions such as *but, although, whereas, while,* that is, those that introduce ideas that contrast, would benefit from a form of pre-processing that would convert them either to individual claims or to two contrasted comparisons prior to their being analyzed using methods described in this study. As for the differences between journals, the *Diabetes* journal achieved 0.74 precision and 0.69 recall. Breast Cancer, *Endocrinology*, and *Carcinogenesis* generally have lower average precision and recall (the average precision ranges from 0.55 to 0.62 and average recall from 0.49 to 0.53). This finding indicates that particular areas and disciplines write differently and that some are clearer than the others.

Finally, for building the model in the last stage in this dissertation, a set of 225 annotated sentences was used that will be made available to the research community so that this method may be further improved in the future. Appendix A in this dissertation lists the examples of direct comparative sentences that have been referenced throughout this work. A few of these sentences from the training set but most of them are not part of the training set. Also, sentences that do not satisfy direct comparison sentences requirements are listed in Appendix B of this document. The sentences included in the Appendices are meant to assist the computational community with understanding the nature and characteristics of direct comparative sentences and how they are different from or similar to other types of comparisons sentences.

To conclude, facets that are extracted from comparative sentences under this dissertation: (1) provide the information required to create of a comparative, multi-document summary; (2) enable one or more entities to be compared over time and/or across collections; (3) track the endpoint over time and/or across collections; (4) track the direct comparative relations over time and/or across collections; (5) identify gaps that exist with relation to comparative studies; (6) identify contradictions that exist between studies—for example, by focusing on the results of the comparisons and establishing where the conclusions are different while the study parameters are the same or similar; and (7) support exploratory search.

REFERENCES

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, 17-21.

Baker, Collin F., Fillmore, Charles J., & Lowe, John B. (1998). *The Berkeley FrameNet Project*. Paper presented at the Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, Montreal, Quebec, Canada.

Berners-Lee, Tim, Hendler, James & Lassila, Ora (2001). The Semantic Web. Scientific American. 284(5): 34-43.

Blake, Catherine. (2003) Information Synthesis: A Mixed-Initiative Meta-Analytic Approach to Facilitate Knowledge Discovery from Scientific Text, Ph.D. Dissertation *Information and Computer Science*, Specialization: Information Access and Management, University of California, Irvine.

Blake, Catherine. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics, 43*(2), 173-189. doi: http://dx.doi.org/10.1016/j.jbi.2009.11.001.

Blake, C., & Lucic, A. (2015). Automatic endpoint detection to support the systematic review process. *J Biomed Inform, 56*, 42-56. doi: 10.1016/j.jbi.2015.05.004

Blake, Catherine & Rindflesch, Thomas, C. (Under review). Leveraging syntax to better capture the semantics of elliptical coordinated compound noun phrases. *Journal of Biomedical Informatics*.

Bloom, Benjamin S. (1956). Taxonomy of educational objectives; the classification of educational goals *(*1st ed.*)*. New York*: Longmans, Green.*

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, E. and et al. (2014). "Findings of the 2014 Workshop on Statistical Machine Translation," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, pp. 12-58.

Bresnan, J.W. (1973). "Syntax of the Comparative Clause Construction in English," *Linguistic Inquiry*, vol. 4, pp. 275-343.

Dai, M., Shah, N., Xuan, W., Musen, M., Watson, S., Athey, B., and Meng, F. (2008). An efficient solution for mapping free text to ontology terms. AMIA Summit on Translational Bioinformatics.

Damerow, J. (2014). *A quadruple-based text analysis system for history and philosophy of science* (Order No. 3631837). Available from ProQuest Dissertations & Theses Full Text; ProQuest Dissertations & Theses Global. (1566477529). Retrieved from https://search.proquest.com/docview/1566477529?accountid=14553.

Donegan, Sarah, Williamson, Paula, Gamble, Carrol, & Tudur-Smith, Catrin. (2010). Indirect Comparisons: A Review of Reporting and Methodological Quality. *PLoS ONE, 5*(11), e11054. doi: 10.1371/journal.pone.0011054.

Finkel, Jenny Rose, Grenager, Trond, & Manning, Christopher. (2005). *Incorporating non-local information into information extraction systems by Gibbs sampling*. Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan.

Friedman, Carole. (1989). "A General Computational Treatment Of The Comparative," presented at the Association of Computational Linguistics, Stoudsburg, PA.

Friedman, Carole. (2000). A broad-coverage natural language processing system. *Proceedings of the AMIA Symposium*, 270-274.

Friedman, Carol, Shagina, Lyudmila, Lussier, Yves, & Hripcsak, George. (2004). Automated Encoding of Clinical Documents Based on Natural Language Processing. *Journal of the American Medical Informatics Association : JAMIA, 11*(5), 392-402. doi: 10.1197/jamia.M1552.

Frunza, Oana, Inkpen, Diana, & Tran, Thomas. (2011). A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts. *IEEE Trans. on Knowl. and Data Eng., 23*(6), 801-814. doi: 10.1109/tkde.2010.152.

Girju, Roxana, Badulescu, Adriana, & Moldovan, Dan. (2006). Automatic Discovery of Part-Whole Relations. *Comput. Linguist., 32*(1), 83-135. doi: 10.1162/coli.2006.32.1.83.

Groza, T., Hassanzadeh, H., & Hunter, J. (2013). Recognizing scientific artifacts in biomedical literature. *Biomed Inform Insights, 6*, 15-27. doi: 10.4137/bii.s11572.

Guichard, M., D’Andon, A., Rumeau Pichon, C., & Borget, I. (2015). PRM209 - The Use of Indirect Comparisons in Medicines Evaluation for their Access To Reimbursement by the Has. *Value in Health, 18*(7), A719. doi: http://dx.doi.org/10.1016/j.jval.2015.09.2724.

Halteren, Hans van, & Teufel, Simone. (2003). *Examining the consensus between human summaries: initial experiments with factoid analysis*. Paper presented at the Proceedings of the HLT-NAACL 03 on Text summarization workshop - Volume 5.

Harris, Zellig. S & Mattick, P., Jr. (1988). Scientific Sublanguages and the Prospects for a Global Language of Science. *Annals of the American Association of Philosophy and Social Sciences* No.495.73-83.

Harris, Zellig S. (2002). The structure of science information. *Journal of Biomedical Informatics, 35*(4), 215-221. doi: http://dx.doi.org/10.1016/S1532-0464(03)00011-X.

Hearst, Marti A. (1992). *Automatic acquisition of hyponyms from large text corpora*. Paper presented at the Proceedings of the 14th conference on Computational linguistics - Volume 2, Nantes, France.

Jindal, N., & Liu, B. (2006). "Identifying Comparative Sentences in Text Documents," presented at the Special Interest Group in Information Retrieval (SIGIR), Seattle Washington USA.

Jonquet, Clement, Shah, Nigam H., & Musen, Mark A. (2009). The Open Biomedical Annotator. *Summit on Translational Bioinformatics, 2009*, 56-60.

Klein, D. & Manning. (2003). "Fast Exact Inference with a Factored Model for Natural Language Parsing. ," in *Advances in Neural Information Processing Systems*, pp. 3-10.

Leonhard, Annette. (2009). *Towards retrieving relevant information for answering clinical comparison questions*. Paper presented at the Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Boulder, Colorado.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., & Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics, 28*(7), 991-1000. doi: 10.1093/bioinformatics/bts071.

Lin, Chin-Yew, & Hovy, Eduard. (2002). *Manual and automatic evaluation of summaries*. Paper presented at the Proceedings of the ACL-02 Workshop on Automatic Summarization - Volume 4, Philadelphia, Pennsylvania.

Lin, Chin-Yew, & Hovy, Eduard. (2003). *Automatic evaluation of summaries using N-gram co-occurrence statistics*. Paper presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, Edmonton, Canada.

Lucic, Ana, & Blake, Catherine L. (2016). Improving Endpoint Detection to Support Automated Systematic Reviews. *AMIA Annual Symposium Proceedings, 2016*, 1900-1909.

Machamer, Peter, Darden, Lindley, & Craver, Carl F. (2000). Thinking about Mechanisms. *Philosophy of Science, 67*(1), 1-25. doi: doi:10.1086/392759.

Marneffe, Marie-Catherine de, & Manning, Christopher D. (2008). *The Stanford typed dependencies representation*. Paper presented at the Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, Manchester, United Kingdom.

Manning, Christopher D., Raghavan, Prabhakar, Schütze, Hinrich. (2008). *Introduction to Information Retrieval*: Cambridge University Press.

Medical Subject Headings Fact Sheet. https://www.nlm.nih.gov/pubs/factsheets/mesh.html (Accessed November, 2, 2016).

Mintz, Mike, Bills, Steven, Snow, Rion, & Jurafsky, Dan. (2009). *Distant supervision for relation extraction without labeled data*. Paper presented at the Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2, Suntec, Singapore.

Moldovan, Dan, Girju, Roxana, & Rus, Vasile. (2000). *Domain-specific knowledge acquisition from text*. Paper presented at the Proceedings of the sixth conference on Applied natural language processing, Seattle, Washington.

Knight, Kevin, & Marcu, Daniel. (2000). *Statistics-Based Summarization - Step One: Sentence Compression*. Paper presented at the Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence.

Krauthammer, Michael, & Nenadic, Goran. (2004). Term identification in the biomedical literature.

Journal of Biomedical Informatics, 37(6), 512-526. doi: http://dx.doi.org/10.1016/j.jbi.2004.08.004.

Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, Pancoast E, Schein A, Ungar L, White P, Winters S. (2004). Integrated annotation for biomedical information extraction. *Proceedings of Biolink 2004.*

McCray, Alexa T., Burgun, Anita, & Bodenreider, Olivier. (2001). Aggregating UMLS Semantic Types for Reducing Conceptual Complexity. *Studies in health technology and informatics, 84*(0 1), 216-220.

Meyers, A., R.Reeves, C.Macleod, R.Szekely, V.Zielinska, B.Young, and R.Grishman.2004.The NomBank Project: An interim report.In Proceedings of the HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation, pages 24–31, Boston, MA.

Musen, Mark A., Noy, Natalya F., Shah, Nigam H., Whetzel, Patricia L., Chute, Christopher G., Story, Margaret-Anne, . . . and the, Ncbo team. (2012). The National Center for Biomedical Ontology. *Journal of the American Medical Informatics Association : JAMIA, 19*(2), 190-195. doi: 10.1136/amiajnl-2011-000523.

Nenkova, Ani, Passonneau, Rebecca, & McKeown, Kathleen. (2007). The Pyramid Method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process., 4*(2), 4. doi: 10.1145/1233912.1233913.

Niculae, V. and Yaneva, V. 2013. Computational considerations of comparisons and similes. In: *Proceedings of ACL 2013 Student Research Workshop*, pp. 89-95. Sofia, Bulgaria, August 2013.

Noy, Natalya F., Shah, Nigam H., Whetzel, Patricia L., Dai, Benjamin, Dorf, Michael, Griffith, Nicholas, . . . Musen, Mark A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research, 37*(Web Server issue), W170-W173. doi: 10.1093/nar/gkp440.

ODM documentation. Oracle Data Mining Administrator’s Guide, 11g Release 2. All Oracle documentation is available online at Oracle Technology Network (http://www.ont.oracle.com).

Palmer, Martha, Gildea, Daniel, & Kingsbury, Paul. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Comput. Linguist., 31*(1), 71-106. doi: 10.1162/0891201053630264.

Park, D. Hoon, & Blake, C. (2012). "Identifying comparative sentences in full-text scientific articles," presented at the Workshop on Detecting Structure in Scholarly Discourse at Association of Computational Linguistics, Jeju, South Korea.

Pratt, Wanda, & Yetisgen-Yildiz, Meliha. (2003). A Study of Biomedical Concept Identification: MetaMap vs. People. *AMIA Annual Symposium Proceedings, 2003,* 529-533.

Rath, G. J., Resnick, A. and Savage, T. R. (1961), The formation of abstracts by the selection of sentences. Part I. Sentence selection by men and machines. Amer. Doc., 12: 139–141. doi:10.1002/asi.5090120210.

Renear, Allen H., & Palmer, Carole L. (2009). Strategic Reading, Ontologies, and the Future of Scientific Publishing. *Science*, *325*(5942), 828-832. doi: 10.1126/science.1157784.

Resnick, A. (1961). Part II. The reliability of people in selecting sentences. *American Documentation, 12*(2), 141-143. doi: 10.1002/asi.5090120211.

Rindflesch, Thomas C., & Fiszman, Marcelo. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics, 36*(6), 462-477. doi: http://doi.org/10.1016/j.jbi.2003.11.003.

Röhl, Johannes. (2012). Mechanisms in biomedical ontology. *Journal of Biomedical Semantics, 3*(Suppl 2), S9-S9. doi: 10.1186/2041-1480-3-S2-S9.

Rosario, Barbara, & Hearst, Marti A. (2004). *Classifying semantic relations in bioscience texts*. Paper presented at the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain.

Salton, Gerard, Singhal, Amit, Mitra, Mandar, & Buckley, Chris. (1997). Automatic text structuring and summarization. *Information Processing & Management*, *33*(2), 193-207. doi: http://dx.doi.org/10.1016/S0306-4573(96)00062-3.

Savova, Guergana K., Masanz, James J., Ogren, Philip V., Zheng, Jiaping, Sohn, Sunghwan, Kipper Schuler, Karin C., & Chute, Christopher G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA, 17*(5), 507-513. doi: 10.1136/jamia.2009.001560.

Shah, Nigam H., Bhatia, Nipun, Jonquet, Clement, Rubin, Daniel, Chiang, Annie P., & Musen, Mark A. (2009). Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics, 10*(Suppl 9), S14-S14. doi: 10.1186/1471-2105-10-S9-S14.

Scheible, Silke. (2007). *Towards a computational treatment of superlatives*. Paper presented at the Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop, Prague, Czech Republic.

Smith, Barry, & Grenon, Pierre. (2004). The Cornucopia of Formal-Ontological Relations. *Dialectica, 58*(3), 279-296. doi: 10.1111/j.1746-8361.2004.tb00305.x.

Song, F., Harvey, I., & Lilford, R. (2008). Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. *J Clin Epidemiol, 61*(5), 455-463. doi: 10.1016/j.jclinepi.2007.06.006

Spärck, Karen, Jones (2007). Automatic summarising: The state of the art. *Inf. Process. Manage., 43*(6), 1449-1481. doi: 10.1016/j.ipm.2007.03.009.

Steinberger, J.; Ježek, K. (2012). Evaluation measures for text summarization. Computing and Informatics, 28: 251–275.

Tanenblatt, M., Coden, A., Sominsky, I. (2010). The ConceptMapper approach to named entity recognition.in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, et al (Eds.) Language Resources and Evaluation. European Language Resources Association, Malta; 2010:546–551.

Teufel, Simone, & Moens, Marc. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics.*, *28*(4), 409-445. doi: 10.1162/089120102762671936.

UMLS basics. Semantic Relationships. https://www.nlm.nih.gov/research/umls/new_users/online_learning/SEM_004.html (Accessed November 2, 2016).

Voorhees, Ellen M. (1998). *Variations in relevance judgments and the measurement of retrieval effectiveness*. Paper presented at the Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia.

von Stechow, A. (1984). Comparing semantic theories of comparison. *Journal of Semantics*. 3, 1–77.

APPENDIX A: TERM DEFINITION AND ANNOTATION INDEX

Direct comparison sentence – To be considered a direct comparison sentence, the sentence needs to reference *at least* two compared entities, the endpoint or the basis on which the entities are compared, and also the result of comparison.

Four *facets* of *direct* comparative sentences:

Entity 1 – First compared entity. Very often, this is the entity that is mentioned first in a comparative sentence. This entity does not appear in the close vicinity to comparison marker. It appears on the opposite end of comparison marker. The numbering is used for easier distinguishing between entities.

Entity 2 – Second compared entity in a comparative sentence. Within the annotation system used in this dissertation, entity 2 is used to indicate the entity that occurs in the vicinity of comparison anchors.

Endpoint – The basis, aspect, property on which the two or more entities are compared.

Relation – Semantic relation that ties the compared entities and the endpoint and indicates whether a property based on which they were compared changed, remained the same, or was different.

Annotation index

[entity 1] – entity 1 (See above)

[entity 2] – entity 2 (See above)

[endpoint] – Basis (property, aspect, endpoint) on which two or more entities in a sentence are compared.

[relation] – Semantic relation that binds the compared entities and the endpoint. Semantic relation indicates whether a property common to compared entities changed, remained the same or was different.

[entity 1_A] – used when there is more than one entity in the sentence that is compared to another entity/ies. For example, if two entities are compared to a third entity, the two entities that are compared to a third entity will be annotated as [entity 1_A] and [entity 1_B] and the entity that they are both compared to (if there is only one) as [entity 2].

[entity 1_AB] or **[entity 2_AB] –** used when the candidate noun refers to and/or qualifies the meaning of both entities: 1_A and entity 1_B.

[relation_A] – used when there is more than one semantic relation that ties the compared entities and endpoints in a sentence.

[relation modifier] – used to indicate a relation modifier such as, for example, *not* or *significantly*. Several relation modifiers can modify only one relation and also one modifier can modify several relations. Because of the important role they have with reference to conveying the semantic meaning of the relation, relation modifiers are indicated in the sentences, however, at the moment of writing this dissertation, only the method for identifying negation modifier has been developed and tested. Identification of other types of relation modifiers is left for future work. Connecting and piecing together the integral parts of sentences that contain multiple comparison claims is also left for future work.

If an annotation (for example, **[entity_1A]** or **[entity 1]** or **[endpoint_1A])** is repeated within the same sentence, this indicates that one of the integral parts of direct comparisons consists of more than one noun phrase. Establishing the boundaries of an integral part in cases when it consists of more than one noun phrase is out of scope of this dissertation. When predicting the role of a noun in a sentence, the expectation is that the classifier will predict the *most relevant* or the main parts of the noun phrases that were manually identified to play the role of entity 1, entity 2 and the endpoint, if not all the noun phrases that sometimes constitute one role.

APPENDIX B: LISTING OF ANNOTATED DIRECT COMPARISON SENTENCES[5]

(5) The result showed that ethanol feeding [entity 1] in rats [entity 1] *increased* [relation] c-Jun mRNA level [endpoint], as compared with the non-ethanol-fed rats [entity 2]. 11470752

(6) The plasma insulin concentration [endpoint_A] at 8 weeks of age [endpoint _A] and the pancreatic insulin content [endpoint_B] and the beta-cell mass [endpoint_C] on day 8 [endpoint_BC] and 8 weeks [endpoint_BC] of age [endpoint_BC] in STZ- treated rats [entity 1] *were severely* [relation modifier] *reduced* [relation] compared with those of normal rats [entity 2] ($P < 0.001$). 14988244

(7) Compared with non- [entity 2_A] and irregular tea drinkers [entity 2_B] in particular, we found *reduction* [relation_A] in circulating levels [endpoint_AB] of both estrone (-13%) [endpoint_A] and estradiol (-8%) [endpoint_B] among weekly/daily green tea drinkers [entity 1_A] and *increase* [relation_B] in both estrone (+19%) [endpoint_C] and estradiol (+10%) levels [endpoint_D] among weekly/daily black tea drinkers [entity 1_B]. 15661801

(11) Fasting glucose [endpoint] was *higher* [relation] in diabetic subjects [entity 1] (168.8 55.2 \ mg/dl) than in nondiabetic subjects [entity 2] (93.9 9.6 mg/dl).12882937

(12) Short-term treatment [entity 1] of ovariectomized rats [entity 1] with estradiol [entity_1] plus progesterone [entity 1] caused *significantly* [relation modifier] *decreased* [relation] preoptic Gal-R1 mRNA levels [endpoint] compared with those after treatment [entity 2] with estrogen only [entity 2]. 9751492

(13) Buserelin-stimulated serum testosterone levels [endpoint_A] in male G_11_alpha knockout mice [entity 1_A] was *significantly* [relation modifier] *higher* [relation_A] than in control mice [entity 2_A], although Buserelin stimulated estradiol levels [endpoint_B] in female G_11_alpha knockout mice [entity 1_B] were *lower* [relation_B] than in control mice [entity 2_B]. 9607776

(14) Furthermore, a trend toward a *lower* [relation] hepatic microsomal free cholesterol [endpoint_A] and triglyceride concentrations [endpoint_B] was observed with atorvastatin [entity 1] compared with simvastatin treatment [entity 2]. 28200735

(15) Thus, *similar* [relation] to norepinephrine [entity 2_A] and epinephrine [entity 2_B],

[5] Direct comparative sentences included in Appendix B come from the three TREC Genomics journals used in this dissertation (*Diabetes*, *Carcinogenesis*, and *Endocrinology*). Several of the sentences included in this listing are included the training set. The annotated training set of 225 sentences, used in the main experiment (Chapter 4), will be released and made available to the research community.

dopamine [entity 1] in the presence of IL-1beta [entity 1] *induced* [relation] a synergistic stimulation [endpoint] of IL-6 release [endpoint]. 9927320

(16) In this RBA assay, raloxifene [entity 1] exhibited the *highest* [relation modifier] *affinity* [relation] for ERalpha [endpoint] relative to estradiol [entity 2]. 10579349

(17) A *twofold* [relation modifier] *increase* [relation] in ALT activity [endpoint] was observed in serum from diabetic rats [entity 1] compared with the lean controls [entity 2].15277384

(18) Fasting serum insulin concentrations [endpoint] *decreased* [relation] *significantly* [relation modifier_A] and *similarly* [relation modifier_B] during both rosiglitazone [entity 1] and metformin therapy [entity 2] by 4±1 and 4±2 mU/l, respectively (Fig 1). 15277403

(20) In conclusion, intensive lifestyle intervention [entity 1] *reduced* [relation] levels [endpoint] of nontraditional cardiovascular risk factors [endpoint] both relative to placebo [entity 2_A] and to a lesser degree relative to metformin [entity 2_B]. 15855347

(21) Interestingly, metformin [entity 1] (Fig 2C) also *inhibited* [relation] PTP opening [endpoint] with an efficacy *similar* [relation] to that of CsA [entity 2]. 15983220

(22) Nevertheless, since hepatic [endpoint_A] and renal responses [endpoint_B] observed in our subjects [entity 1] treated with metformin [entity 1] *did not* [relation modifier] *differ* [relation] from those not treated with it[6] [entity 2], an influence of antecedent metformin treatment on our results seems unlikely. 12765948

(23) Although the mechanism [entity 1] by which metformin [entity 1] activates AMPK [endpoint] remains unclear, it must be *different* [relation] from that of AICA riboside [entity 2], which acts by being converted to the AMP mimetic agent, ZMP. 12145153

(24) Among all participants, including those who developed diabetes, fasting glucose [endpoint_A], insulin [endpoint_B], and proinsulin concentrations [endpoint_C] were *significantly* [relation modifier_A] *lower* [relation_A] than placebo [entity 2] at the first annual visit in the metformin [entity 1_A] and the lifestyle groups [entity 1_B] and *increased* [relation_B] during the 2nd and 3rd years, with the levels remaining *significantly* [relation modifier_C] *lower* [relation_C] than in the placebo group [entity 2] (Fig 2). 16046308

[6] Although in this sentence the second compared entity is not mentioned explicitly, this sentence is listed as a direct comparison sentence. Note that entity 2 in this case should be identified as “those not treated with it.” This sentence is an example of a sentence in which an entity is not a noun but a part of a relative clause which is the reason why “those not treated with it” is not underlined.

(25) The small effect [endpoint] of insulin [endpoint] to stimulate Akt activity [endpoint] before metformin treatment [entity 1] was *similar* [relation] to that in the troglitazone group [entity 2] before treatment (NS) [entity 2]. 11812753

(26) However, in contrast to troglitazone treatment [entity 2], there was *no* [relation modifier] *enhancement* [relation] of Akt activation [endpoint] in response [endpoint] to insulin [endpoint] after metformin treatment [entity 1] (Fig 2B). 11812753

(41) Stimulation [endpoint] of the islets [endpoint] with 3 mmol/l glucose [endpoint] did *not* [relation modifier_A] show significant [relation modifier_B] *differences* [relation] between the wild-type [entity 1] and IRS-1 KO groups [entity 2] (data not shown). 15161756

(42) Nonfasting plasma glucose levels [endpoint_A] and the overall glycemic excursion [endpoint_B] (area under the curve) to a glucose load [endpoint_B] were *significantly* [relation modifier] *reduced* [relation] (1.6-fold; $P < 0.05$) in (Pro_3) GIP-treated mice [entity 1] compared with controls [entity 2]. 16046312

(43) There is evidence to suggest that the somatic mutational pathway [endpoint_A] may *differ* [relation_A] between invasive [entity 1] and LMP ovarian tumours [entity 2] and invasive tumours [entity 1] are *more likely* [relation_B] than LMP [entity 2] to exhibit p53 overexpression [endpoint_B]. 11159743

(46) ADX + CORT animals [entity 1] had plasma corticosterone concentrations [endpoint] *similar* [relation] to sham animals [entity 2] (6.3 2.8 microg/dl) ($P > 0.05$). 14633853

(48) Only *modest* [relation modifier_A] *reductions* [relation] *(although significant)* [relation modifier_B] were seen in fibrinogen levels [endpoint] in the lifestyle group [entity 1] relative to the metformin [entity 2_A] and placebo group [entity 2_B]. 15855347

(49) A *58%* [relation modifier] *decrease* [relation] in mammary tumor incidence [endpoint] was demonstrated in DMBA-treated rats [entity 1] fed 20% less food/day [entity 1] when compared with ad libitum-fed carcinogen treated controls [entity 2]. 9934852

(50) B6 LFD mice [entity 1] had body [endpoint_A] and liver weights [endpoint_B] *similar* [relation] to those found in 129 HFD mice [entity 2]. 15855315

(51) In the present study, we confirmed our previous observation that 2 d [entity 1] of fasting [entity 1] in male rhesus monkeys [entity 1] *triggers* [relation] neuroendocrine responses [endpoint] *similar* [relation] to those observed in healthy men [entity 2]. 11796492

(52) Additionally, the maximum inhibition [endpoint_A] in tumor incidence [endpoint_A] and

multiplicity [endpoint_B] with Targretin [entity 1] was *similar* [relation] to that achieved with tamoxifen [entity 2]. 10874003

(53) It is interesting to note that in vitro, the affinity [entity 1_AB] and potency [entity_1_AB] of tamoxifen [entity 1_A] and raloxifene [entity 1_B] on the ER [endpoint] is *similar* [relation] to that of 17beta-estradiol [entity 2]. 9389534

(54) *Similar* [relation] to the effects [entity 2] of troglitazone [entity 2], metformin treatment [entity 1] *significantly* [relation modifier] *decreased* [relation] HbA_1c [endpoint_A], glucose [endpoint__B], and insulin concentrations [endpoint_C]. 11812753

(55) This work also provided evidence that the mechanisms by which benfluorex [entity 1] *reduces* [relation] hepatic gluconeogenesis [endpoint] are *markedly* [relation modifier] *different* [relation] from those of metformin [entity 2], the main antidiabetic compound used in the world 12145146

(56) HbA_1c [endpoint] was *higher* [relation] in diabetic women [entity 1] than men [entity 2] (P = 0.004). 12031985

(57) By contrast, GK fetuses [entity 1] *exhibited* [relation_A] a *h*igher [relation_A] plasma glucose concentration [endpoint_A] and a *lower* [relation_B] plasma insulin level [endpoint_B] (P < 0.001) as compared with values in Wistar fetuses [entity 2]. 11812746

(58) Mean blood pressure [endpoint] was *reduced* [relation] in response to captopril [entity 1_A] (P < 0.001) or candesartan (P < 0.001) therapy [entity 1_B] as compared to untreated SHRs [entity 2]. 11272159

(59) These results suggested that insulin sensitivity [endpoint] was *increased* [relation] in VPAC2R KO mice [entity 1] compared with their WT siblings [entity 2]. 12239111

(60) Compared with placebo [entity 2], fenofibrate [entity 1] *significantly* [relation modifier] *decreased* [relation] plasma concentrations [endpoint_A] of triglycerides [endpoint_A], total apoB [endpoint_B], apoCIII [endpoint_C], and lathosterol [endpoint_D], as well as the VLDL triglyceride-to-apoB [endpoint_E] and lathosterol-to-cholesterol ratios [endpoint_F]. 12606523

(61) The addition [entity 1] of flutamide [entity 1] with testosterone propionate [entity 1] *produced* [relation] a *significant* [relation modifier] *reduction* [relation] in clathrin heavy chain mRNA [endpoint] compared with the effect of supplementation with only testosterone propionate [entity 2] (P < 0.05; n = 5). 9529000

(62) The Av3hGK-treated diabetic mice [entity 1] also *displayed* [relation] a 64.4% [relation

modifier] *reduction* [relation] in fasting plasma insulin levels [endpoint] as compared with control groups [entity 2] at week 1 posttreatment [entity 2]. 11574410

(63) Measurement of plasma insulin levels revealed that Av3hGK treatment [entity 1] *resulted* [relation] in a *significant* [relation modifier] *reduction* [relation] in fasting insulin levels [endpoint] in the diabetic mice [entity 1] (66%) compared with both of the control groups [entity 2] (Fig 2B). 11574410

(64) After caffeine ingestion [entity 1], plasma FFA [endpoint] was *significantly* [relation modifier] *higher* [relation] compared with placebo [entity 2]. 11872654

(65) There *was* a *significant* [relation modifier] (44%) *increase* [relation] in insulin sensitivity [endpoint] compared with baseline [entity 2] after 26 weeks [entity 1] (P < 0.01) in the rosiglitazone group [entity 1]. 12453903

(67) In the PD, hybridization signal intensity [endpoint] is *enhanced* [relation] in the TRH group [entity 1_A] and *significantly* [relation modifier] *reduced* [relation] in the T_4 group [entity 1_B] compared to that in controls [entity 2]. 9048604

(68) Moreover, VOR treatment [entity 1] did not *significantly* [relation modifier] *decrease* [relation] cortical thickness [endpoint] in contrast with orchidectomy [entity 2]. 9165015

(69) Furthermore, after IGF-I treatment, insulin-stimulated tyrosine phosphorylation [endpoint_AB] of the IR [endpoint_A] and IRS-1 [endpoint_B] in muscle [endpoint_AB] of the LID mice [entity 1] (Fig 5C and D) was *comparable* [relation] to that seen in the control mice [entity 2]. 11334415

(70) This may in part explain the *lack* [relation modifier] of a *difference* [relation] in initiation [endpoint] we observed between African American [entity 1] and non-Hispanic White women [entity 2]. 16168103

(71) Eight (57.1%) [entity 1_A] of these patients[7] [entity 1_A] had an anthracycline-free interval [endpoint_A] greater_than_12 months [endpoint_A], nine (64.2%) [entity 1_B] were classified as having non-anthracycline-resistant disease [endpoint_B], and seven (50%) [entity 1_C] were taxane naïve [endpoint_C], resulting in *no remarkable* [relation modifier] *differences* [relation] when compared with the full patient population [entity 2]. 16168103

(72) In hypoglycemic fed rats [entity 1_A], numbers of Fos-immunoreactive neurons [endpoint_A] in

[7] Although this is an anaphoric reference, this sentence is included as an example of a direct comparison sentence.

the LHA [endpoint_A] were *slightly* [relation modifier_A] but *not significantly* [relation modifier_B] *higher* [relation_A] than in vehicle-treated controls [entity 2_A], whereas the hypoglycemic fasted rats [entity 1_B] had *significantly* [relation modifier_C] *more* [relation_B] Fos_+ nuclei [endpoint_B] than both other groups[8] [entity 2_B], with nearly twice as many as in controls [entity 2_B] (P < 0.001). 11147774

[8] Anaphoric reference.

APPENDIX C: LISTING OF ANNOTATED NON DIRECT COMPARISON SENTENCES[9]

(1) We compared control [entity 1] with treated animals [entity 2]. 12538615

(2) Parous rats [entity 1] were compared with respective age-matched virgins (AMVs) [entity 2]. 10223190

(3) We compared arterial pressure [endpoint] between mice [entity 1] on normal [entity 1] or high-fat diet [entity 2]. 15983201

(4) This study compared the frequency of p53 mutations [endpoint] in BRCA1-associated breast carcinomas [entity 1] with that in sporadic breast tumors [entity 2] in a prevalence sample [entity 2] of Ashkenazi Jewish women [entity 2]. 10070948

(8) Anatabine [entity 1_A] and anabasine [entity 1_B] like nicotine [entity 2_A] and cotinine [entity_2_B] are non-carcinogenic. 12082012

(9) Hence, it is possible that the TGF-beta1 growth response is more dependent on the amount of TbetaR-II receptor expression than it is on the TGF-beta1-PRL response. 9681516

(10) This is 262 times lower than the median levels found in Chinese workers. 15817613

(19) Water intake [endpoint] was randomly monitored throughout the study, was found to *increase* [relation] in proportion [endpoint] to body weight [endpoint] (*r* = 0.69, *P* < 0.001), and was *not* [relation modifier] *different* [relation] among treatment groups (0.093 ± 0.004, 0.098 ± 0.011, and 0.098 ± 0.004 ml · g^{-1} · day^{-1} for control [entity 1_A], metformin-[entity 1_B], and rosiglitazone-treated mice [entity 1_C], respectively). 15983227[10]

[9] Although some of the sentences that do not satisfy the requirements of a direct comparative sentence are annotated (i.e., it is indicated which noun can be classified as entity 1 or entity 2 or the endpoint in the sentence), the annotations are included only to highlight the difference and draw attention to the missing elements which, most frequently, are the reason why the sentence does not satisfy the requirements of a direct comparative sentence. Several sentences listed in this Appendix (e.g. sentences 28 and 29) satisfy the requirements of a direct comparative sentence but they communicate the method of the study rather than its result which is the reason why they were not included in the list of direct comparison sentences. However, it should be mentioned that if the models built under this dissertation were to be applied on an unseen collection, the models would not be able to differentiate between the sentences that communicate the method of the study and those that communicate the result of the study.

[10] This sentence features *different among* anchor that was considered but eventually excluded from the set of comparison anchors used in this dissertation.

(27) Hepatic glucose release was calculated as the difference between the systemic glucose release and renal glucose release. 12765948

(28) Hyperlipidemia was present in 50% and hypertension in 70% of the patients, without differences between the groups. 15448095

(29) The cumulative incidence of diabetes in BB.7_b animals is similar to that observed in BBDP rats originating from the BRM colony (data not shown). 12351436

(30) It has been recognized for >15 years that methylation patterns in tumor cells are altered relative to those of normal cells. 10688866

(31) There are considerable data to support the concept that the type of fat or fiber is actually more important to tumor development than is the amount of either of these components in the diet. 10910952

(32) Note that the amount of 35_S-labeling was similar in all lanes, whereas 125_I-iodine incorporation was increased with LPO. 16037381

(33) Because of the differences between inulin and insulin itself, whether delivery of the bioactive hormone is increased remains speculative. 15504953

(34) To what extent the cell dispersion procedure promotes a biological situation similar to inflammation or injury is unclear. 9348205

(35) Conversely, at higher (nanomolar range) concentrations the effects of melatonin resulted in a stimulation of GH secretion, in marked contrast with the cAMP levels, which continued to decrease. 12960030

(36) However, FM1-43 was reported to more likely stain the lipid membrane by unrestricted lateral diffusion through the membrane than by aqueous diffusion , and therefore, the kinetics of the rising phase of FM1-43 fluorescence might be similar between these exocytic processes. 16123364

(37) In men, a single injection of a relatively large dose of rhFSH (3000 IU) resulted in less than a 2-fold increase in inhibin B levels. 12639898

(38) Differences between groups were analyzed by the Student's t test. 12663468

(39) Total RNA from two control mice (nos 2 and 3) with weight gain similar to the DCA-treated mice were used as controls. 11470764

(42) There were no differences among treatment groups in the rate of discontinuation or the reasons for discontinuation. 15983221

(44) One other major difference between these two strains was identified in this study. 11522683

(45) SERCA2a mRNA levels were compared with nondiabetic levels on a Northern blot. 11916940

(47) In two recent studies with type 2 diabetic patients [entity 1], the effects of the PPARgamma agonist troglitazone [entity 1] on insulin signaling [endpoint_A] and action [endpoint_B] were therefore compared with the effects of metformin [entity 2]. 12196460

(66) The mean hepatic IGF-I mRNA abundance [endpoint] (IGF-Ia transcript) was *reduced* [relation] by 70% [relation modifier_A] compared with ad libitum-fed controls [entity 2] and by *50%* [relation modifier_B] compared with pair-fed controls [entity 2] 9048593[11]

[11] This sentence does not include a reference to entity 1.