The Design of Lexical Database for Indonesian Language

D Gunawan[1*] and A Amalia[2]

[1]Department of Information Technology, Faculty of Computer Science and Information Technology University of Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155

[2]Department of Computer Science, Faculty of Computer Science and Information Technology University of Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155

[*]danigunawan@usu.ac.id

Abstract. Kamus Besar Bahasa Indonesia (KBBI), an official dictionary for Indonesian language, provides lists of words with their meaning. The online version can be accessed via Internet network. Another online dictionary is Kateglo. KBBI online and Kateglo only provides an interface for human. A machine cannot retrieve data from the dictionary easily without using advanced techniques. Whereas, lexical of words is required in research or application development which related to natural language processing, text mining, information retrieval or sentiment analysis. To address this requirement, we need to build a lexical database which provides well-defined structured information about words. A well-known lexical database is WordNet, which provides the relation among words in English. This paper proposes the design of a lexical database for Indonesian language based on the combination of KBBI 4th edition, Kateglo and WordNet structure. Knowledge representation by utilizing semantic networks depict the relation among words and provide the new structure of lexical database for Indonesian language. The result of this design can be used as the foundation to build the lexical database for Indonesian language.

1. Introduction

Research related to natural language processing, text-mining, information retrieval or sentiment analysis require a lexical database. The example of widely used lexical database for English is WordNet [1]. WordNet is manually build by a group of lexicographers in Princeton University. The result is compiled as lexical database files. However, WordNet is also available as Resource Description Framework [2] which reuses several well-known ontologies. There are many researchers try to build WordNet from other languages. For example, Korean WordNet [3] as well as Romanian WordNet [4] were built automatically using existing WordNet [5] and available lexical resources. Some researchers also try to build WordNet for Indonesian language. A prototype version of Indonesian WordNet synsets is built with Monolingual Lexical Resources [6]. The author limit on synsets building only. Another research builds Indonesian WordNet by mapping existing WordNet entries to Indonesian word sense definition [7].

Instead of mapping existing WordNet to Indonesian word sense, our approach is combining existing sources of Indonesian online dictionaries, which are KBBI 4th edition (we will mention this as KBBI) and Kateglo with WordNet. Because the structure of Indonesian language is different with

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

English. By combining KBBI 4th edition, Kateglo and WordNet, we propose a new design structure of lexical database for Indonesian language. This design might provide the basic structure to build the database model or the ontology for lexical database for Indonesian language. Previous publication suggests to reuse the other well-known ontology if it satisfy our requirements [8].

This paper is organized as follows. In section 2, we describe the challenge in building lexical database for Indonesian language. In section 3, we explain the approach to design a new structure of lexical database for Indonesian language. Finally, we draw some conclusions and future works in section 4.

2. Challenges in Building Lexical Database for Indonesian Language

Building a lexical database for Indonesian Language is a challenging task. First of all, the lexical database for Indonesian language requires a robust structure as the house of data. Although there is a robust structure of lexical database in English, the structure requires modification before it is applied to another lexical database. The different nature of the way a language is organized to the others is a reason why the lexical database structure cannot be applied directly. We may use the basic structure available to build a lexical database for Indonesian language. However, it requires modification before we apply the structure for Indonesian language.

Second, the usage of a lexical database for Indonesian language. The usage of a lexical database is mainly for machine, not human. KBBI and Kateglo are the samples of online dictionaries that can be a good reference for human. However, machine cannot read the meaning of words from KBBI or Kateglo easily. To enable the machine for reading the meaning of words easily, the lexical database should expose all properties of a word.

Third, designing a lexical database for machine means the data should be ready to follow the structure. The data for lexical database for Indonesian language are obtained from several different sources. This requires special treatment for each data sources.

3. The Proposed Design of Lexical Database for Indonesian Language

The structure of WordNet, KBBI and Kateglo have their own advantages. As the official dictionary of Indonesian Language and the availability of data is from KBBI, we start building the structure from KBBI point of view. We combine WordNet and Kateglo advantages to support the structure from KBBI. The proposed design structure of lexical database for Indonesian language consists of six main categories. These categories are Types, Styles, Etymology, Part-of-Speech, Disciplines and Lexical Domain.

Mostly, the categorization in KBBI is also used in Kateglo. Because the main source of Kateglo is KBBI. Besides KBBI, there are a few sources of Kateglo such as Badan Bahasa, Bahtera, Daisy Subakti, Sofia Mansoor and Wikipedia. Although the main source of Kateglo is KBBI, it does not fully apply the KBBI structure. Kateglo has slightly different point of view with the KBBI. Kateglo has data that will enrich the lexical database for Indonesian Language. For example, it has more entries especially in abbreviation and acronym. Kateglo have more than 3000 entries of abbreviations and acronyms. Besides, Kateglo also has entries about the misspelled words, proverbs and glossaries.

Generally, the structure of the online version of KBBI 4th edition is not as complex as WordNet. The KBBI consists of several categories. However, we cannot directly map the KBBI contents to the WordNet structure. Those are Kelas Kata (Part-of-speech), Bidang (Discipline), Jenis (Type), Asal Kata (Etymology), and Ragam (Style). The approach to organize words in KBBI is different with WordNet. Some terms in KBBI structures may not exist in WordNet and vice versa.

Both KBBI and WordNet have part-of-speech category. However, the properties are slightly different. KBBI has seven properties, meanwhile WordNet has six properties. We observe that some of both properties shares the same meaning and some others have their own meaning. As shown in table 1, sub-categories in KBBI such as Numeralia, Partikel and Pronomina do not have pairs of properties in WordNet. Conversely, Adjective Satellite and Phrase in WordNet properties do not have pairs in KBBI. Meanwhile, Kateglo does not fully follow the properties of part-of-speech category in

KBBI. Kateglo has the same sub-categories of part-of-speech in general. However, it extracts a property (Partikel) to be more detail categorization. They are Preposisi, Konjungsi, Interjeksi, Bentuk Terikat, Partikel lain.

Table 1. Part-of-speech between KBBI and WordNet

KBBI	WordNet
Adjektiva	Adjective
-	Adjective Satellite
Adverbia	Adverb
Nomina	Noun
Numeralia	-
Partikel	-
Pronomina	-
Verba	Verb
-	Phrase

We eliminate all the properties from WordNet that have no pairs with KBBI. Moreover, we add several part-of-speech properties from Kateglo to enrich word classification. The part-of-speech properties from Kateglo have more details properties for Partikel. After combining part-of-speech properties in Kateglo, the proposed Part-of-speech is shown in table 2.

Table 2. The combination of part-of-speech between KBBI and Kateglo

Part-of-speech
Adjektiva
Adverbia
Nomina
Numeralia
Partikel
Pronomina
Verba
Preposisi
Interjektori
Konjungsi
Bentuk terikat
Partikel lain

Originally, category Jenis (Types) consists of properties Kata Dasar (Root Word), Kata Berimbuhan (Word Affixes), Gabungan Kata (Phrase), Kiasan (Idiom) and Peribahasa (Proverb). After observing the Kateglo and social media, there are types of words that can be included in Types category. Those are Singkatan Kata (Abbreviation) and Akronim (Acronym). Abbreviation is often used in informal short messages such as social media, short message service (SMS), email and so on.

Meanwhile, acronym is often used in both formal and informal text. As shown in figure 2, the sentence consists of common abbreviation and acronym. Jokowi is an abbreviation which stands for Joko Widodo. Besides, HUT and TNI are the acronym that stand for Hari Ulang Tahun and Tentara Negara Indonesia respectively. The proposed new style for designing lexical database is shown in table 3.

Table 3. Proposed new types properties

Jenis (Types)
Kata dasar (Root word)
Kata berimbuhan (Word affixes)
Gabungan kata (Phrase)
Kiasan (Idiom)
Peribahasa (proverb)
Singkatan kata (abbreviation)
Akronim (Acronym)

Some terms from local tribes' languages or foreign languages which are used frequently by the Indonesian people are considered to be included in KBBI. Starting from the 4th edition, KBBI is now support another new category, which is etymology. Therefore, this category might be used to explore the origin of a word. However, etymology is not supported yet by the WordNet structure. Asal Kata (Etymology) category is used by KBBI and Kateglo. As WordNet does not have etymology in its structure, there is no modification needed to Asal Kata category. There are many properties in Asal Kata (Etymology) such as Melayu, Inggris, Padang, and so on.

Another category in KBBI is Ragam (Style). There are some properties such as Arkais (Archaic), Cakapan (Conversation), Hormat (Polite), Kasar (Impolite) and Klasik (Classic). Arkais is used to label the words that almost never been used anymore. While the words in sub-category Cakapan are used in conversation. The words that are used to express politeness is categorized in Hormat. Conversely, the words in property Kasar expresses impoliteness. Property Klasik is used to categorize the old words. We cannot find the similar category in WordNet. Similar to category Jenis (Types), we modify category Ragam (Styles) from KBBI according to the observation of Kateglo and social media. Kateglo provides the data about misspelled words along with the correct words. According to this feature, we add a new properties called Salah Eja (Misspelled) to category Ragam (Styles). For example, akte instead of akta, artifak instead of artefak, and so on. In addition, we also add informal word that usually used by the people called Slang. For example, words "bokap", "beneran", "woles" are informal words that substitute "ayah", "benarkah", "santai" respectively.

Table 4. Proposed new types properties.

Ragam (Styles)
Arkais (Archaic)
Cakapan (Conversation)
Hormat (Polite)
Klasik (Classic)
Kasar (Impolite)
Salah Eja (misspelled)
Slang (Slang)

Discipline is a new category in KBBI 4th edition. Previous edition does not include discipline to categorize the words. The words in KBBI are also categorized based on the disciplines. For example, the words laman (homepage), peramban (browser), unduh (download), daring (online) are included in teknologi informasi (information technology) discipline. Other disciplines such as Elektronika (Electronics), Biologi (Biology), Kedokteran (Medicine), Ekonomi (Economics) and so forth have their own words. By referring to discipline category, people might be aware to use the words for special purpose sentences. In addition, machine might also take the advantages of discipline categories. For example, automatic keywords extraction for certain discipline might refer to dictionary to validate the selected words in related disciplines.

Both KBBI and Kateglo do not support Lexical Domain. Lexical Domain is supported by WordNet. For example, a word "Tiger" will recognized as noun only if it is searched with KBBI or Kateglo. However, WordNet has deeper result that expose the word "Tiger" is a noun and the lexical domain is noun animal. According to this feature, we include lexical domain to enrich lexical database for Indonesian language.

4. Conclusion and Future Research

This paper explores combination of KBBI, Kateglo and WordNet to design a new structure of lexical database for Indonesian language. We use this approach based on our assumption that the organization of Indonesian language is different with other languages. Basically we utilize existing features on KBBI and Kateglo as our main sources of Indonesian words to build a lexical database. However, WordNet as a widely used lexical database has more feature that can be adapted to our design. For the future research, we are going to implement the design to build the ontology. Then we utilize semantic web to build the lexical database. We hope that the design of lexical database for Indonesian language can be reference to other language similar to Indonesian language to build another lexical database.

5. Acknowledgments

The authors gratefully acknowledge that the present research is supported by Ministry of Research and Technology and Higher Education Republic of Indonesia. The support is under the research grant TALENTA USU of Year 2016 Contract Number 91/UN5.2.3.1/PPM/SP/2016.

References

[1] Miller A G 1995 WordNet: A Lexical Database for English *Commun. ACM* **38** pp 39-41

[2] Princeton University 2016 WordNet RDF Princeton University *Available: http://wordnet-rdf.princeton.edu*

[3] Lee C, Lee G and Yun S J 2000 Automatic WordNet mapping using word sense disambiguation *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*

[4] Barbu E and Mititelu V B 2007 Automatic Building of Wordnets *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005* **292**

[5] Farreres X, Rigau G and Rodffguez H 1998 Using WordNet for Building WordNets *Proc. of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*

[6] Gunawan and Saputra A 2010 Building Synsets for Indonesian WordNet with Monolingual Lexical Resources *2010 Int. Conf. on Asian Language Processing*

[7] Putra D D, Arfan A and Manurung R 2008 Building an Indonesian WordNet *Proceedings of the 2nd International MALINDO Workshop*

[8] Gunawan D 2015 Protocol for E-Commerce data harvesting *Technology, Informatics, Management, Engineering & Environment (TIME-E), 2015 International Conference on*