SANDIA REPORT
SAND2008-6104
Unlimited Release
Printed September 2008

Identification of Threats Using Linguistics-Based Knowledge Extraction

Peter A Chew

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2008-6104
Unlimited Release
Printed September 2008

Identification of Threats Using Linguistics-Based Knowledge Extraction

Peter A Chew
Cognitive and Exploratory Systems Department
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185

ABSTRACT

One of the challenges increasingly facing intelligence analysts, along with professionals in many other fields, is the vast amount of data which needs to be reviewed and converted into meaningful information, and ultimately into rational, wise decisions by policy makers. The advent of the world wide web (WWW) has magnified this challenge. A key hypothesis which has guided us is that threats come from ideas (or ideology), and ideas are almost always put into writing before the threats materialize. While in the past the 'writing' might have taken the form of pamphlets or books, today's medium of choice is the WWW, precisely because it is a decentralized, flexible, and low-cost method of reaching a wide audience. However, a factor which complicates matters for the analyst is that material published on the WWW may be in any of a large number of languages.

In 'Identification of Threats Using Linguistics-Based Knowledge Extraction', we have sought to use Latent Semantic Analysis (LSA) and other similar text analysis techniques to map documents from the WWW, *in whatever language they were originally written*, to a common language-independent vector-based representation. This then opens up a number of possibilities. First, similar documents can be found across language boundaries. Secondly, a set of documents in multiple languages can be visualized in a graphical representation. These alone offer potentially useful tools and capabilities to the intelligence analyst whose knowledge of foreign languages may be limited. Finally, we can test the over-arching hypothesis – that ideology, and more specifically ideology which represents a threat, can be detected solely from the words which express the ideology – by using the vector-based representation of documents to predict additional features (such as the ideology) within a framework based on supervised learning.

In this report, we present the results of a three-year project of the same name. We believe these results clearly demonstrate the general feasibility of an approach such as that outlined above. Nevertheless, there are obstacles which must still be overcome, relating primarily to how 'ideology' should be defined. We discuss these and point to possible solutions.

This page is intentionally left blank.

1 Executive Summary

'Identification of Threats Using Linguistics-Based Knowledge Extraction', or ITUL for short, is an approach we have developed over the course of a three-year research project to assist in analysis of documents in any of 54 languages (this number could easily be increased with the availability of more text data). Machine translation is the usual approach to dealing with documents in unknown languages, but it is subject to error and can result in unreliable results 'downstream'. Our approach completely bypasses the need for machine translation, taking documents in their original language and converting them directly into a numerical vector-based representation (in a 'language-independent semantic space') which can then be used for multiple purposes, including creating visualizations of document sets, mining for features such as 'hostile ideology', and so on. All these applications are intended to make life easier for the analyst burdened by an overwhelming volume of data to explore. Tests of the system indicate that it has a high level of accuracy (for example, in multilingual document clustering we obtain a precision of around 90%). Using ITUL, we have also obtained promising results in automated classification of documents according to whether they express a hostile ideology or not. There are likely many other uses for such a system; in general, it would be useful in any situation where a large number of multilingual documents have to be processed or analyzed, and when some of the languages may not be familiar to the analyst.

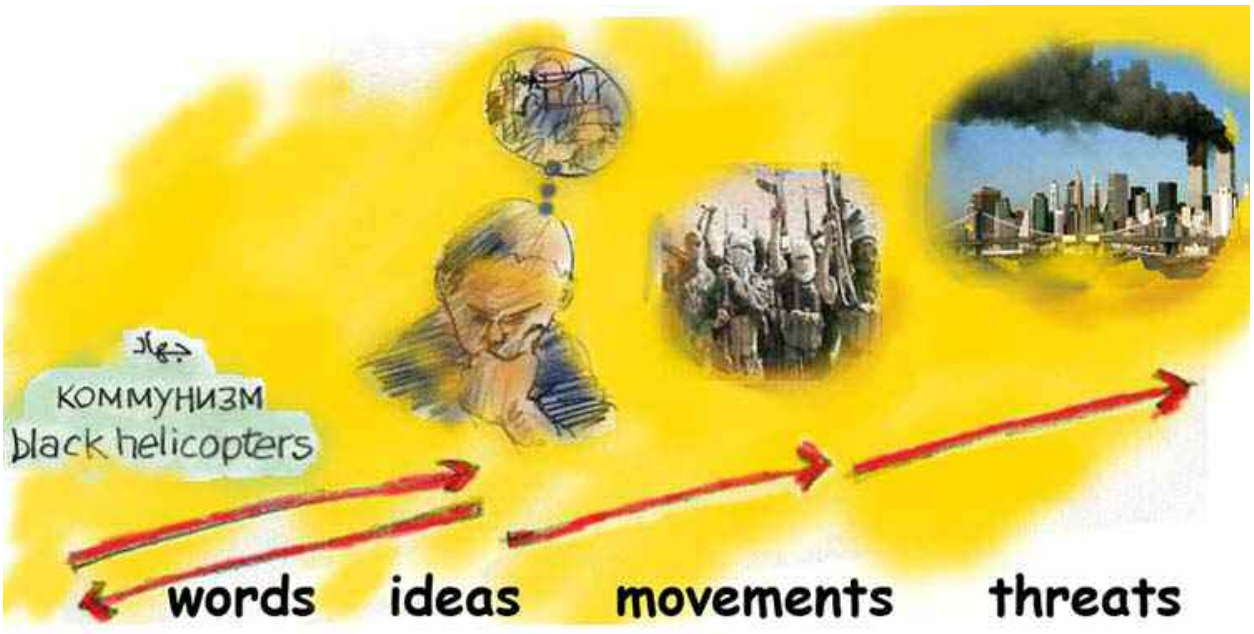

Figure 1. The relationship between words, ideology, and threats

2 Introduction

This report is intentionally brief, since much of the research in the 'Identification of Threats' project has already been documented elsewhere. Where appropriate, we refer the reader to our related publications.

The basic hypothesis we explore is that ideology is somehow contained within words, or text. This idea is depicted in Figure 1. Historical examples supporting this hypothesis are Hitler's *Mein Kampf* and Lenin's *Что делать?* (*What is to be Done?*). It is clear just from these examples that ideology can be expressed in any language, and that an approach which is limited to one or just a few languages is, by definition, very incomplete. Because of this, an essential part of this project was to construct a framework, ITUL, to allow documents to be analyzed *no matter what their original language was*. The construction of ITUL is described in Section 3, and its usefulness is not limited to the analysis of threats or ideology; it can be used to solve a wide range of problems involving multilingual text.

In Section 4, however, we return to the problem of determining a document's ideology based on its text. We refer to results which show that ITUL appears to be promising as one component in a framework for identifying ideology in text.

In our mind, however, there are still some questions as to how ideology should be defined, and we examine these in Section 5. While the usefulness of ITUL for a variety of purposes is without question for us, the lack of a clear definition of 'ideology' leads us to exercise caution in assessing the value of this framework (or any other based on text) for identifying ideology, hostility, or threats in text.

3 ITUL: a framework for analysis of multilingual text

When dealing with text in more than one language, the key question we are interested in is the following: how do we abstract away from the *words*, which are language-specific and which might be regarded as 'noise', to the underlying meaning or concepts? If this question can be answered satisfactorily, and if the association between text and concepts can be represented in a principled (for example numeric or vector-based) fashion, then the door is opened to compare text in different languages computationally.

Much of our work in the 'Identification of Threats' project has been directed at solving this problem. Early on, we took the decision to use the parallel text of the Bible as a 'Rosetta Stone' from which, essentially, to reverse-engineer existing translation. The Bible has the advantages that it is the world's most translated book (partial translations in over 2,400 languages, full translations in over 400 languages, and electronically-available translations in close to 100 languages); it can be aligned by verse; and translations are generally public-domain.

Approaches we attempted were as follows:

(1) Without linear-algebraic decomposition: Chew et al. (2006)

(2) The 'textbook' approach to multilingual text analysis, Latent Semantic Analysis (LSA) applied to a parallel corpus: Chew and Abdelali (2007)

(3) PARAFAC2: Chew et al. (2007)

(4) Latent Morpho-Semantic Analysis (LMSA): Chew et al. (2008a). This approach involves breaking words down into morphemes (their minimal units of meaning, such as 'read+ing'), then applying LSA. The approach has greatest value for languages which rely on complex morphology, such as Arabic and Russian.

(5) Latent Semantic Analysis with Term Alignments (LSATA): Bader and Chew (2008). This approach applies eigenvalue decomposition to a term-alignment matrix, leveraging key insights from Statistical Machine Translation within a vector-space model.

In each case, we evaluated performance by measuring precision, treating chapters in the Quran (also translated into a number of languages) as queries and looking at how frequently the chapters' counterparts in different languages were correctly retrieved. Full discussion of the evaluation metric is given in, for example, Chew et al. (2008b), Bader and Chew (2008), and Chew et al. (2008c).

Overall, the best results were obtained using Latent Morpho-Semantic Analysis with Term Alignments, or LMSATA, which is a combination of LMSA and LSATA. Full details of LMSATA are given in Chew et al. (2008c). To summarize, LMSATA correctly identified translations in 93.12% of cases we tested, and achieved a precision of 88.18% in multilingual clustering. This was a significant improvement on our initial results, where a precision of just 26.44% was obtained in multilingual clustering.

Perhaps the clearest way to convey what we ultimately achieved, however, is impressionistic. In Figures 2 and 3 below, we show a visualization of the books of the Bible in 5 languages using Tamale (visualization software created at Sandia by Tamara Kolda and Ann Yoshimura). Generally, documents which appear close to one another in the graph are those which are most semantically similar, although it must be borne in mind that the visualization attempts to represent a 300-dimensional semantic space in two dimensions, so the visualization is inevitably an approximation. It is clear from Figure 3 that books are generally most similar to their counterparts in other languages – the lowest-level clusters – and then they cluster into similar topics. Thus, for example, Matthew, Mark, Luke, John and Acts (all of which are on similar topics) are all shown in one sub-area of the graph in Figure 2. While the dataset here is artificial, in the sense that documents do not always have translations in the real world, one can imagine that a similar visualization of documents from the WWW (for example) would allow an analyst quickly to see which documents are on similar topics, even without knowing the language of the documents and even if the documents are in different languages.

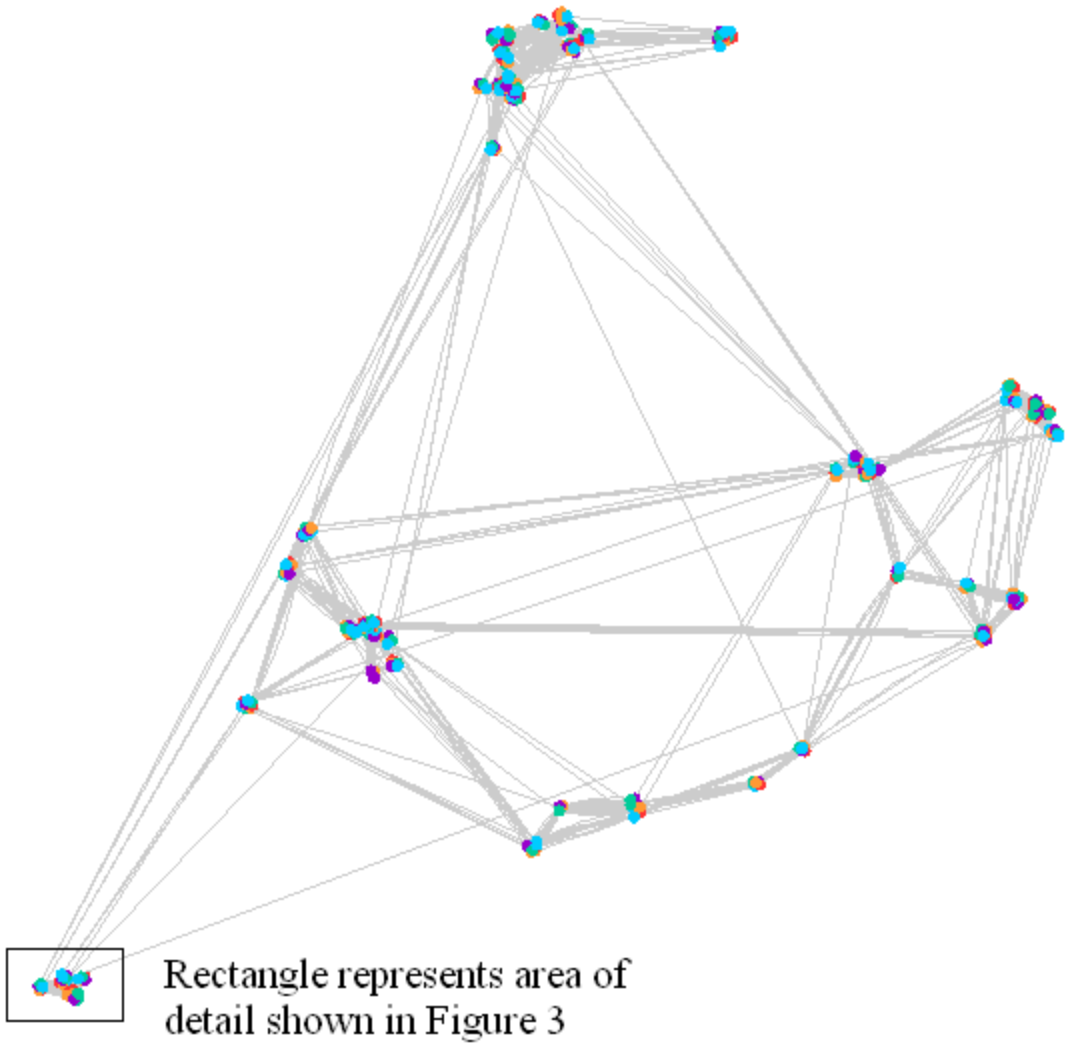

Figure 2. Visualization of multilingual Bible books in vector space

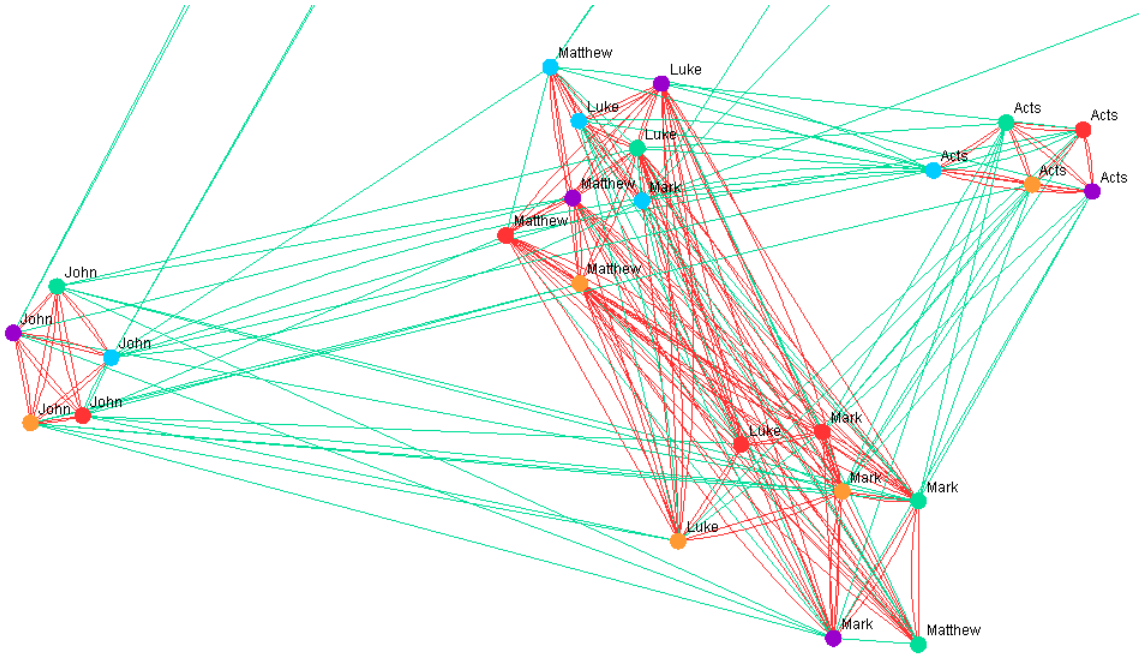

Figure 3. Partial visualization of multilingual Bible books in vector space

4 Determining ideology from text

The ITUL framework referred to above, and described in detail in Chew et al. (2008c), treats each document as a 'bag of words' in its original language, then uses the parallel corpus (the 'Rosetta Stone') to convert this directly to a vector representation. These vectors are then used to calculate similarities, a necessary step both for evaluating precision and for creating visualizations of the sort shown in Figures 2 and 3.

The vectors can also be used as input for supervised learning, as follows. Suppose there is a sample of documents for which we know some other feature, e.g. its author or the ideology it expresses. For a much larger number of documents, perhaps, we do not know the correct classification. There are a large number of machine learning algorithms, such as neural networks, support vector machines, decision trees, Naïve Bayes, and so on, which allow a classifier to be trained on the examples where the correct classification is known (the training set), and then deployed on the other cases (the deployment set).

In our case, as mentioned, the ITUL framework associates every document with a vector based on the words contained in the document. This is true both for training and deployment documents. In machine learning, each entry in the vector can be treated as an independent variable. The feature we want to predict, such as ideology, is then the dependent variable. The nature of the relationship between the independent and dependent variables need not be known; if there is *some* relationship, a machine learning algorithm should be able to discern it and use it to make predictions about new cases. Put another way, we should be able to train the algorithm on documents of known ideology, and if ideology truly is contained in text, then we should be able to use the algorithm to make predictions about the ideology of new documents.

Our attempts to do just this are documented in Chew et al. (2008a). The results we obtained appear to indicate that our approach to predicting ideology is promising: we correctly distinguish between Lenin's and Hitler's writings with an accuracy of 94.7% (with a baseline of 65.1%), and between hostile and non-hostile writings with an accuracy of 98.9% (with a baseline of 90.0%). This is evidence in support of the hypothesis that ideology *is* contained in text alone, and moreover can be automatically extracted using techniques such as ITUL.

5 The path forward

While the results we obtained for ideological classification are promising, we think that some caution needs to be exercised over their interpretation. It is known that authors employ distinctive writing styles (which are often manifested in the choices of high-frequency words). The same is true in speech, where different speakers gravitate towards different 'filler' words such as 'you know' or 'really'. This facilitates the task of classifying documents by author, and may mean, for example, that the 94.7% accuracy we obtained in classifying Lenin versus Hitler is in fact the accuracy of author classification rather than ideological classification.

However, this in turn raises the question of how to separate the author from the ideology. Is it possible to speak of Leninism without Lenin, or German National Socialism without Hitler? On some level, it would seem that it should be, since there are Leninists and Nazis today, even though both Lenin and Hitler died decades ago. Yet one might also question whether these ideologies would exist today if it had not been for the historical individuals who formulated them in words.

If it is possible to separate the author from the ideology, then it may be necessary to formulate more clearly how ideology and language are related. It seems that taking this step would open the way to an objective, clearer evaluation of the feasibility of an approach such as the one we have developed for ideological classification. We are actively seeking to continue this research path, and will do so as further funding permits.

The utility of ITUL, the framework for multilingual text analysis, is however not dependent on the success of our experiments with ideological classification. We believe that ITUL stands by itself as a success of this project, and anticipate its use in real-world applications involving large-scale multilingual text.

6 Acknowledgement

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

7 References

Brett W. Bader and Peter A. Chew. 2008 (forthcoming). Enhancing Multilingual Latent Semantic Analysis with Term Alignment Information. *Proceedings of COLING 2008*.

Peter A. Chew, Steve J. Verzi, Travis L. Bauer and Jonathan T. McClain. 2006. Evaluation of the Bible as a Resource for Cross-Language Information Retrieval. *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, 68-74.

Peter A. Chew and Ahmed Abdelali. 2007. Benefits of the 'Massively Parallel Rosetta Stone': Cross-Language Information Retrieval with over 30 Languages. *Proceedings of the Association for Computational Linguistics,* 872-879.

Peter A. Chew, Brett W. Bader, Tamara G. Kolda, and Ahmed Abdelali. 2007. Cross-Language Information Retrieval Using PARAFAC2. *Proceedings of the Conference on Knowledge Discovery and Data Mining*, 143-152.

Peter A. Chew, W. Philip Kegelmeyer, Brett W. Bader and Ahmed Abdelali. 2008. The Knowledge of Good and Evil: Multilingual Ideology Classification with PARAFAC2 and Machine Learning. *Language Forum* 34 (1), 37-52.

Peter A. Chew, Brett W. Bader, and Ahmed Abdelali. 2008 (forthcoming). Latent Morpho-Semantic Analysis: Multilingual Information Retrieval with Character N-Grams and Mutual Information. *Proceedings of COLING 2008*.

Peter A. Chew, Brett W. Bader, Ahmed Abdelali, Stephen Helmreich, and Steve J. Verzi. Forthcoming. *Can Computational Linguistics Improve Information Retrieval?*

DISTRIBUTION:

1 MS 0123 Donna Chavez, LDRD Office, 1011
1 MS 0899 Technical Library, 9536
5 MS 1012 Peter Chew, 6343


Sandia National Laboratories