Text Mining and Natural Language Processing – Introduction for the Special Issue

Anne Kao & Steve Poteet
Boeing Phantom Works
P.O. Box 3707, MC 7L-43
Seattle, WA 98124
anne.kao@boeing.com,
stephen.r.poteet@boeing.com

ABSTRACT

This paper provides an introduction to this special issue of SIGKDD Explorations devoted to Natural Language Processing and Text Mining.

Keywords

Text Mining, Natural Language Processing, Text Analysis.

1. INTRODUCTION

There is a wide range of technologies and focus areas in Human Language Technology (HLT). These include areas such as Natural Language Processing (NLP), Speech Recognition, Machine Translation, Text Generation and Text Mining. In this issue, we will focus on two of these areas: NLP and Text Mining.

NLP has been around for a number of decades. It has developed various techniques that are typically linguistically inspired, i.e. text is typically syntactically parsed using information from a formal grammar and a lexicon, the resulting information is then interpreted semantically and used to extract information about what was said. NLP may be deep (parsing every part of every sentence and attempting to account semantically for every part) or shallow (parsing only certain passages or phrases within sentences or producing only limited semantic analysis), and may even use statistical means to disambiguate word senses or multiple parses of the same sentence. It tends to focus on one document or piece of text at a time and be rather computationally expensive. It includes techniques like word stemming (removing suffixes) or a related technique, lemmatization (replacing an inflected word with its base form), multiword phrase grouping, synonym normalization, part-of-speech (POS) tagging (elaborations on noun, verb, preposition etc.), word-sense disambiguation, anaphora resolution (who does "he" or "the CEO" refer to), and role determination (e.g. subject and object).

Text Mining is more recent, and uses techniques primarily developed in the fields of information retrieval, statistics, and machine learning. Its aim typically is not to understand all or even a large part of what a given speaker/writer has said, but rather to extract patterns across a large number of documents. The simplest form of Text Mining could be considered information retrieval, also called text retrieval or document retrieval, what typical search engines do. However, more properly Text Mining consists of areas such as automatic text classification according to some fixed set of categories, text clustering, automatic summarization, extraction of topics from texts or groups of text and the analysis of topic trends in text streams. While information retrieval and other forms of text mining frequently make use of word stemming, more sophisticated techniques from NLP have been rarely used.

We organized a panel discussion on the interaction between NLP and Text Mining in SIGKDD 2004. This special issue attempts to further investigate work in these two areas. It includes nine papers from six countries and regions covering a range of application areas including web mining, bioinformatics, criminal investigation and analysis of computing log files.

2. INFORMATION EXTRACTION

The following group of articles explores information extraction, mostly named entity extraction, largely using machine learning techniques rather than hand-built rules, with a varying amount of linguistic information being used.

Mooney and Bunescu do a wonderful job summarizing various contributions machine learning has made in advancing the information extraction task. They examine a number of information extraction algorithms, both as methods for extracting useful knowledge in and of themselves and as precursors to (or integrated parts of) various data mining algorithms. Although they look at a great variety of algorithms and applications, the most linguistic knowledge any of them use is POS tagging.

Gliozzo et al. provide a discussion on how to use Instance Filtering to reduce the size of the training set for supervised classification-based learning systems for entity recognition. Their work shows that not only can this improve the computation time for both training and classification, but, by reducing the skewness in the data, accuracy can improve as well. They also use POS, but no other linguistic concepts.

Fu et al. explore the usefulness of two extensions to current Hidden Markov Models for entity extraction in Chinese: including using the preceding literals as well as tags to predict entities (lexicalization) and grouping characters into known words as input to entity extraction. Grouping characters into words is a long standing problem in any text analysis task in Chinese (as well as in Japanese and Korean) and could be considered an NLP technique akin to phrasal grouping in languages like English. However, their experiments show that, while lexicalization helps with entity extraction in Chinese, incorporating known word information does not.

While the focus of Li et al. is on a new distributed form of association rule mining, they use the results of their entity extraction algorithm which is discussed briefly. It is another example of a machine learning algorithm for finding entities in text and, like some of those above, uses POS tagging and no other

NLP techniques. They use an example of narrative text of police investigation reports to illustrate their new method.

3. CLASSIFICATION

There are four papers that concern themselves essentially with text classification. The first two use no NLP techniques at all beyond word stemming or lemmatization.

Liu et al. discuss the effects of taking advantage of hierachical structure for text classification using Support Vector Machines (SVMs) based on a large scale taxonomy, like the Yahoo! Directory. They conclude that, while the hierarchical approach results in enormous speed-up of training and classification, the poor quality of the classification due to the skewed distribution of the Yahoo! Directory and other large taxonomies is not improved at all. This is consistent with our experience in running large scale classification applications in Boeing (with from 500 classes to 60,000 classes).

Peng et al. show that incorporating temporal information into text classification, either via a modified Naïve Bayes algorithm or a Hidden Markov Model, can improve the classification of computer system log files.

The next two papers exploit considerably more NLP techniques, including parsing, in the service of text classification.

Mullen et al. present some preliminary results of using machine learning (specifically text classification) to identify rhetorical zones in a scientific article (e.g. whether the specific stretch of text is relating background or discussing methods or results). The purpose of rhetorical zone identification is to help with the interpretation of the results of information extraction, although that is not the focus of the paper. They explore Naïve Bayes and SVMs to perform the text classification, but, in addition to lemmatized words and word bi-grams, they also use grammatical roles derived from an NLP parser.

Popowich provides an overview of a system that uses a combination of NLP techniques and text mining to score insurance claims on several categories. He uses a statistical POS tagger, a partial parser, abbreviation and acronym expansion, named entity extraction, the lexical resources of WordNet, and role determination to perform an initial analysis of text, with a taxonomy of concepts and hand-built rules to extract the concepts from the text, and finally a weighted mapping of the concepts into the insurance categories of interest. Note that the extraction of concepts is much like information extraction and it is in the service of text classification. He reports high recall and precision for the system, but is unable to compare it to a more traditional text classifier that did not make use of as much NLP technology.

4. SEARCH, AND MINING OF SEARCH RESULTS

Liang et al. was the only paper to directly address the problem of how NLP might benefit text mining in general and search in particular, following co-author Marchisio's position in SIGKDD 2005's panel discussion. They describe their system, InFact, which uses NLP and a unique representation to index free text for rapid search based on richer primitives than traditional search engines allow and subsequent mining of the results. They claim that NLP allows them to achieve a higher degree of knowledge discovery. Unfortunately, while they show that the user can construct what seem to be more refined queries, they did not present any experimental evidence to demonstrate quantitative improvements. Of course, given the fact that their queries are different than standard queries, it might be difficult to construct the appropriate experiments.

5. FUTURE WORK

This collection of papers shows a rather wide range of applications using various text mining and NLP techniques. What the field needs now is a sober scientific assessment of what linguistic concepts and NLP techniques are beneficial for what text mining applications. This would involve a clear classification of the various linguistic concepts that might be of use (e.g. part-of-speech, grammatical role, phrasal parsing) and the various technologies for getting at these concepts (e.g. full parsing vs. shallow parsing vs. heuristics to get at role information), as well as a classification of text mining applications and of properties of text and corpora (collections of text data). It would further involve innovative experimental designs and new approaches to evaluation. Finally, it would require some hard work comparing various techniques on a wide range of application types and corpora.

With the growing needs in Bioinformatics and in web mining, further research in these areas should definitely benefit a lot of applications. We hope our collection in this issue can serve as a catalyst to motivate more research work in this area.

6. ACKNOWLEDGMENTS

The authors wish to thank our colleagues in Boeing Phantom Works, Mathematics and Computing Technology for providing reviews and comments for the paper submissions: William Ferng, Dragos Margineantu, Winson Taam, Rod Tjoelker, Changzhou Wang, and Jason Wu. They represent a wide range of expertise, including machine learning, data mining, statistics, database and mathematics, as a complement to the authors' background in Text Mining and NLP.

About the authors:

Anne Kao is an Associate Technical Fellow in Boeing Phantom Works, Mathematics and Computing Technology. She has been leading NLP projects since 1990 in Boeing, and started Text Mining work in Boeing in 1995. She and her team invented TRUST (Text Mining Using Subspace Representation) which is the text mining engine inside Starlight (a 3-D information Visualization tool by Pacific Northwest National Lab). She has a M.A and Ph.D. in Philosophy, specializing in Philosophy of Language. She also holds an M.S. in Computer Science.

Steve Poteet is an Advanced Computing Technologist in Boeing Phantom Works, Mathematics and Computing Technology. With advanced degrees in linguistics and cognitive science, he has both led and contributed to NLP and text mining projects at Boeing since 1990.