Report on KDD Conference 2004 Panel Discussion Can Natural Language Processing Help Text Mining?

Anne Kao
Boeing Phantom Works
P.O. Box 3707, MC 3707
Seattle, WA 98124
anne.kao@boeing.com

Steve Poteet
Boeing Phantom Works
P.O. Box 3707, MC 3707
Seattle, WA 98124
stephen.r.poteet@boeing.com

ABSTRACT

With large amounts of text data now available on-line, both on the Internet and in corporate repositories, text mining is an area of growing interest. Historically, Data Mining researchers have come out of the statistics, database and machine learning communities. Despite a few exceptions, it has by and large had little interaction with the computational linguistics and natural language processing (NLP) community. Given that the subject matter of text mining is free text, one might naturally assume techniques developed over the last few decades in computational linguistics and NLP, should make big contributions towards the younger field of text mining. However, other than in the area of information extraction, empirical evidence has not borne this out.

Keywords

Text Mining, Natural Language Processing, Text Analysis.

1. INTRODUCTION

Natural Language Processing (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 or shallow, and 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, multiword phrase grouping, synonym normalization, anaphora resolution, and role determination.

Text Mining is more recent, and uses techniques primarily developed in statistics and machine learning. It's 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. It includes things like text classification according to some fixed set of categories, automatic text clustering, extraction of topics from texts or groups of text and the analysis of trends.

In this panel, we discussed

(1) Can traditional NLP methods help text mining? If so, can they help all areas of text mining or just some areas? Which NLP areas/techniques are useful?

(2) What is novel about text mining vs. NLP? In light of this, what would be some new future directions for NLP in light of requirements from text mining?

The discussion was very well attended; there were 200-300 people there. The chair has received a lot of comments during the preparation of the panel discussion from several experts in this area who were unable to participate and there was a lot of interest exhibited by many researchers.

In order to avoid a non-productive debate on what "Text Mining" should cover, we stipulated that we would take a broad view on the areas and applications that Text Mining deals with.

2. PANELISTS

There were five panelists:

- Ken Church (Microsoft Research)
- Oren Etzioni (University of Washington)
- Marko Grobelnik (Jozef Stefan Institute, Slovenia)
- Dave Lewis (David Lewis Consulting)
- Giovanni Marchisio (Insightful)

Panel Chair: Anne Kao (Boeing Phantom Works)

Moderators:

- Anne Kao (Boeing Phantom Works)
- Steve Poteet (Boeing Phantom Works)

3. SUMMARY OF POSITIONS

Marchisio is the only panelist who holds fast to the position that NLP definitely helps the quality of Text Mining, based on personal experience. Lewis holds a fairly skeptical view that NLP helps Text Mining, because no one has produced a scientific study to prove that point. The rest of the panelists were prepared to argue either position, and did just that to keep the panel discussion lively and interesting.

3.1 The range of Language Processing Technology

Lewis pointed out that Language Processing Technology broadly includes the following application areas:

- Text Retrieval (e.g. Search Engine)[1]
- Text Classification (e.g. categorization, filtering, alerting)
- Summarization / automated abstracting
- Question Answering
- Information Extraction
- Machine Translation

[1] For the purpose of this paper, we will use Text Retrieval and Information Retrieval (IR) interchangeably.

The first two are generally recognized as part of Information Retrieval, and the last two as NLP, with the middle two being claimed by both fields.

The goals of Language Processing Technology are:

- *Canonicalization:* Map texts with similar meanings to same data value. This enables counting, conditional actions, and discovering patterns.
- *Condensation:* Replace texts with more compact representations carrying similar meaning. This speeds human or automated processing
- Lewis further made the point that text data can be either 1st person or 3rd person based. With 1st person based data, the text as a whole is directly linked to the entity (e.g. a product complaint is associated with the person making the complaint), whereas 3rd person based data "asserts things about entities". He claims for 1st person data, IR is usually sufficient, but for 3rd person data, NLP may be needed. In general, if IR methods are sufficient, the we should stick to them and only resort to NLP methods when necessary.

He concluded that there is no single right answer to the choice between IR and NLP for text mining, but rather that the choice should always pay attention to the task one is performing.

Grobelnik suggested that text processing should be looked at in terms of levels, including:

- Character level
- Word level (e.g. stemming)
- Shallow tags (part-of-speech, phrases, named entities)
- Sentence level (syntax, logical form)
- Inter-sentential (coreference, discourse)

The first three constitute shallow representations and the last two rich representations. He pointed out that while there are clearly applications that require rich representations (e.g. machine translation, ontology construction, some question answering), there are far more applications that can do well with shallow representations. Incorporating NLP into Text Mining is further restricted by getting representations usable by Text Mining from NLP and the unavailability of good parsers.

3.2 NLP does not help Text Mining

Church looked at one example of where, intuitively, NLP should help, word sense disambiguation. By determining which sense a word refers to in a document, it should help improve the accuracy of IR. Church used this example and pointed out that this thesis has been demonstrated to be incorrect. [1] Unless automatic word disambiguation has a high level of accuracy, it actually results in poorer results in IR.

Church further pointed out the power of data driven approach, whether it is called machine learning, or empiricism.

3.3 NLP helps Text Mining

Marchisio was the loner who held firmly to the position that NLP helps Text Mining. He thinks that we need to be more creative in the way we preprocess and parameterize text, and invent a new indexing standard. An example of synergy between NLP and Text Mining is the use of syntactic roles in IE, Relationship Extraction, IR, and Text Categorization. He proposed a new indexing standard which will bring NLP and Text Mining together. In this proposed scheme, there are four levels of processing: parsing, linguistic normalization (e.g. anaphora resolution), indexing (which supports dynamic analysis), and finally text mining applications. He used examples from his own work to show empirically the gains that can be made by incorporating NLP into Text Mining.

3.4 What can Text Mining do for NLP?

Church proposed that a "Rising tide of data lifts all boats" and quoted his colleague Eric Brill's only somewhat tongue-in-cheek thesis that "It never pays to think until you've run out of data". Given the immense amount of data that has only recently become available on the world-wide web, all approaches to text processing, from more statistical or machine learning approaches to knowledge rich approaches like NLP will benefit.

Following along these lines, Etzioni raised this interesting question to close the loop, after The web provides so much data, that simple approaches can actually extract a lot of knowledge, which should in turn help NLP overcome one large roadblock, the construction of extensive knowledge bases. Examples that he cited include: spell checking, co-location statistics, and semantic category information.

4. FUTURE WORK

People who have less background in NLP, intuitively assume that the use of NLP should definitely improve the quality of Text Mining. However, most current results do not support this position. The panelists pointed out that there is a range of applications types that Text Mining covers. The closer an application is to Document Retrieval, the less NLP seems to help. The closer it is to Information Extraction, the more NLP seems to contribute. There has not been any scientific evidence that can demonstrate the use of say noun phrase" identification can improve the recall-precision of Document Retrieval. On the other hand, statistical methods which pay little attention to sentence structures generally have a harder time with Information Extraction tasks. There is a big need to produce more measurable results than anecdotal theses on how NLP can help Text Mining.

The panelists further pointed out, an even more important question is what can Text Mining do to help NLP, with the development of various new machine learning techniques and with the extremely large amount of web data available. The hope of using Text Mining to help NLP reach a highly scalable level leads to promises that NLP can further benefit Text Mining in the future.

The authors will guest edit an issue in SIGKDD Explorations in the first half of 2005 to further explore this issue/premise/thesis.

5. ACKNOWLEDGMENTS

We wish to thank all the panelists for their participation, as well as the help and encouragement from our team in the Boeing Phantom Works, Mathematics and Computing Technology.

6. REFERENCES

[1] Sanderson, Mark. "Word Sense Disambiguation and Information Retrieval". *SIGIR-94*. 1994. http://dis.shef.ac.uk/mark/cv/publications/papers/my_papers/SIGIR94.pdf