Extracting Significant Words from Corpora for Ontology Extraction

Dileep Damle
KMi, The Open University.
Milton Keynes, MK7 6AA, UK
d.g.damle@open.ac.uk

Victoria Uren
KMi, The Open University.
Milton Keynes, MK7 6AA, UK
v.s.uren@open.ac.uk

ABSTRACT
We show a new method for term extraction from a domain relevant corpus using natural language processing for the purposes of semi-automatic ontology learning. Literature shows that topical words occur in bursts. We find that the ranking of extracted terms is insensitive to the choice of population model, but calculating frequencies relative to the burst size rather than the document length in words yields significantly different results.

Categories and Subject Descriptors
I.2.7 Natural Language Processing – text analysis

General Terms
Algorithms, Experimentation

Keywords
Ontology Learning, Term Extraction

INTRODUCTION
This work is part of a larger project to build ontologies semi-automatically by processing a collection of domain relevant documents. Ontologies are useful in information retrieval and web navigation applications. Manual ontology construction is slow and expensive requiring scarce expertise in the domain and in ontology engineering. Although existing resources such as text, glossaries, terminologies, data models, ontologies for parts of the domain are all useful for ontology construction (See Maedche[10]), often only text is available.

We aim to extract ontologies from natural language text about a given domain. The process is to identify important terms in the domain, identify significant fragments containing the terms and interpret these to yield concepts, and their properties and relations.

Here, we focus on term identification, usually the first step in most methods for ontology learning. Section 2 outlines the previous work which uses parametric population models of word frequency and raises the question which motivates this work Section 3 describes our new method which uses the word frequencies relative to the burst length in contrast with the usual use word frequencies relative to document length. Section 4 presents and discusses some interesting results. Section 5 describes future work.

K-CAP'05, October 2–5, 2005, Banff, Alberta, Canada.
ACM 1-59593-163-5/05/0010.

PREVIOUS RESEARCH
Within corpus semi-automatic methods for term extraction typically involve estimating the parameters of models of distributions of relative frequencies of words over the documents in a corpus. Other methods compare the domain corpus with a general language corpus to identify words with significantly different domain behavior. We take this latter approach, as it also deals with low frequency terms.

Domain corpora are usually small yielding quite low word frequencies and care is required to ensure the statistical validity of the parametric models used. In the literature, Dunning [6]criticized Church et al[1], Gale & Church [7, 8] for lack of statistical validity and also Dumais et al. [5] and Schvaneveldt [13] for not addressing the statistical validity of their internal processing. Yet, the promotion of Poisson and Poisson mixtures continues (see Church [3] and Church and Gale [2], Katz[9]). This has continued till quite recently. Navigli and Velardi[11] use no parametric statistics, Drouin [4] uses normal populations.

The base statistic in all these methods is the is within document relative word frequency. We compared the term rankings yielded by both the Poisson and Binomial models using this definition of relative frequency and found no real differences for any coarse part of speech over 3 different corpora. The condition, $np(1-p)>5$, necessary for the Poisson approximation to Binomial to be valid was not met for any content-words in our data. So, we reviewed the definition of the relative frequency statistic.

As topical words are bursty, it seemed unreasonable to calculate word frequencies relative the document length in words. It is more reasonable to use word frequencies relative to the length of the burst in which the words occur..

THE METHOD
Although we obtained identical results from 3 quite different corpora, in this abstract we only show the results for nouns from the e-book– 'Neuroscience of psychoactive substance use and dependence'[12]. We treated the chap-

ters as separate documents and counted lemmas within parts of speech (e.g. substance/N, consume/V etc.).

We ignored all non-content words. The general language corpus we used was a Natural & Pure Science subset of the BNC – world edition. Both the aggregation over lemmas and discounting of non-content words increases the relative frequencies.

We carried out 3 experiments for each domain corpus:

a. Using Poisson Model and document as window

b. Using Binomial Model and document as window

c. Using Poisson Model with burst as window

An individual burst was considered to end when the lemma did not recur for 50 consecutive non-content lemmas. We used the log likelihood ratios and chi-square tests to test the significance of differences (see [6]). The term lists were ranked on the basis of the chi-square value.

RESULTS AND DISCUSSION

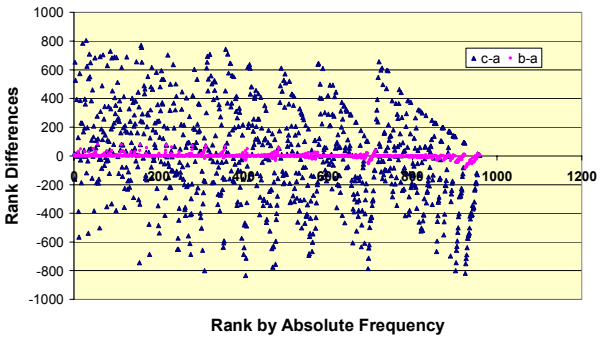

Figure 1. Rank Differences for Nouns

For each part of speech, we ranked lemmas by decreasing chi-square value for each lemma l for the three experiments. The reordering by the experiments b and c relative to the experiment was calculated as rank(a)-rank(b) and rank(a)-rank(c). These two sets of rank differences for nouns are plotted in figure 1 against rank by total number of occurrences. Experiments a and b give identical results but c gives quite a different ordering. Our new method is quite different. Unfortunately, there is no objective evaluation method reported in the literature for term extraction and we are arranging a subjective evaluation using domain experts.

FUTURE WORK

We intend to add Named Entity Recognition, Anaphora Resolution and Word Sense Disambiguation into our term extraction process and experiment with WordNet senses instead of lemmas. This should take us closer to concepts which we need for ontology construction. This work will run in parallel with construction of the ontology extractor using a connectionist approach.

ACKNOWLEDGEMENTS

This work was carried out with support from the UK EPSRC and the Open University.

REFERENCES

1. Church, K.W., et al. *Parsing, word associations and typical predicate-argument relations*. in *International Workshop on Parsing Technologies, CMU*. 1989.
2. Church, K.W. and W.A. Gale, *Poisson Mixtures.* Journal of Natural Language Engineering, 1995. **1**(2): p. 163-190.
3. Church, K.W. *Empirical estimates of adaptation: the chance of two Noriegas is closer to p/2 than p2*. in *17th International Conference On Computational Linguistics*. 2000..
4. Drouin, P., *Term Extraction using non-technical corpora as a point of leverage.* Terminology, 2003. **9**(1): p. 99-115.
5. Dumais, S., et al. *Using latent semantic analysis to improve access to textual information*. in*, CHI '88*. 1988.
6. Dunning, T., *Accurate Methods for the Statistics of Surprise and Coincidence.* Computational Linguistics, 1994. **19**(1): p. 61-74.
7. Gale, W.A. and K.W. Church. *Identifying word correspondence in parallel texts*. in *Proceedings of a workshop on Speech and natural language*. 1991. Pacific Grove, California, United States: Morgan Kaufmann Publishers Inc.
8. Gale, W.A. and K.W. Church, *A program for aligning sentences in bilingual corpora.* Computational Linguistics, 1993. **19**(1).
9. Katz, S.M., *Distribution of content words and phrases in text and language modelling.* Natural Language Engineering, 1996. **2**(1): p. 15-59.
10. Maedche, A. and S. Staab, *Ontology learning for the Semantic Web.* Intelligent Systems, IEEE [see also IEEE Expert], 2001. **16**(2): p. 72-79.
11. Navigli, R. and P. Velardi, *Learning Domain Ontologies from Document Warehouses and Dedicated Web sites.* Computational Linguistics, 2004. **30**(2).
12. Neuroscience of psychoactive substance use and dependence. 2004, WHO: Geneva.
13. Schvaneveldt, R., Pathfinder Associative Networks: Studies in KnowledgeOrganization. 1990, Norwood, NJ: Ablex.