The Effect of Word Sense Disambiguation Accuracy on Literature Based Discovery

Judita Preiss and Mark Stevenson
Advanced Computing Research Center
Department of Computer Science
The University of Sheffield
211 Portobello, Sheffield S1 4DP
United Kingdom
j.preiss@sheffield.ac.uk
mark.stevenson@sheffield.ac.uk

Keywords
Data mining, literature based discovery, LBD, biomedical text processing, word sense disambiguation, WSD

1. INTRODUCTION
The volume of published research in the biomedical domain has required researchers to focus attention to their own area. Consequently, connections between information published in different areas may be missed. Literature based discovery (LBD) attempts to address this issue by searching for terms that connect unrelated pieces of knowledge. For example, assume there is no known connection between disease A and treatment C but that one publication states that disease A affects B and another that treatment C affects B. LBD would hypothesise a connection, between A and C, and that it is worthy of further exploration.

Lexical ambiguity is prevalent within biomedical literature. For example, possible meanings of the term cold include common cold, cold sensation and Chronic Obstructive Lung Disease. However, lexical ambiguity leads to LBD over-generation since the LBD system hypothesises connections via multiple senses of ambiguous words.

We explore the connection between WSD accuracy and the effectiveness of LBD. We examine the effect of a number of WSD approaches with significantly different performance on the hidden knowledge generated and evaluate using a time-slicing approach [2]. We find that applying WSD leads to a significant reduction in LBD over-generation and that LBD is sensitive to WSD accuracy.

2. METHOD
Our LBD system [1] builds a matrix $A$ in which element $a_{ij}$ describes the frequency with which one concept, $c_i$, is linked by title co-occurrence to another concept, $c_j$, in the document collection. A combination of matrix multiplication and subtraction reveals hidden knowledge, i.e. for any concept, the system will generate all terms that can be reached via any single linking term.

We compare three different WSD systems, a general personalized page rank based system which we apply to the biomedical domain, a vector space model based WSD system applicable to any domain but tuned to biomedical texts, and MetaMap which is designed for biomedical publications, and a random baseline.

3. RESULTS
An evaluation is carried out using timeslicing [2], where hidden knowledge is generated from articles published prior to a certain cutoff date and a gold standard is extracted from publications after the cutoff date (any 'knowledge' appearing in publications after the cutoff is deemed new). Results show the performance of the LBD system to be correlated with the performance of the WSD system employed.

4. CONCLUSION
This paper explores how the problem of lexical ambiguity affects the performance of LBD systems and the extent to which WSD could be applied to solve this issue. WSD approaches with varying levels of accuracy were combined with an LBD system. Evaluation of the hidden knowledge generated was carried out using the time-slicing approach and revealed that LBD is sensitive to the accuracy of the underlying WSD system. We therefore conclude that WSD forms a useful component of LBD systems and suggest that further improvements in WSD accuracy could benefit LBD.

5. ACKNOWLEDGMENTS
The work described here was funded by the Engineering and Physical Sciences Research Council (EP/J008427/1).

6. REFERENCES
[1] J. Preiss, M. Stevenson, and R. Gaizauskas. Exploring relation types for literature-based discovery. Journal of the American Society for Information Science and Technology, 2015.
[2] M. Yetisgen-Yildiz and W. Pratt. A new evaluation methodology for literature-based discovery. Journal of Biomedical Informatics, 42(4):633–643, 2009.