

Towards Semantic Literature Based Discovery

Judita Preiss and Mark Stevenson and M. Heidi McClure

Department of Computer Science,
Sheffield University,
Regent Court, 211 Portobello,
Sheffield, S1 4DP
United Kingdom

{J.Preiss, R.M.Stevenson, M.H.McClure}@sheffield.ac.uk

Abstract

Previous systems for literature based discovery, an automatic method of identifying hidden knowledge, have largely been based on bag of words approaches which perform only limited semantic analysis and interpretation. We describe the shortcomings of these approaches and suggest possible solutions that make use of techniques from Natural Language Processing.

Background

The aim of literature based discovery (LBD) is to analyse research literature to identify hidden knowledge. LBD systems identify hidden knowledge by searching for concepts which are connected, with two main design approaches: **open** discovery, starting from a given concept, A , extracts connected phrases to form a set B , each of whose elements are further branched to connected phrases, C . One of the elements of C is manually selected to be the target concept. Alternatively, **closed** discovery starts with known concepts A and C , a set of connected concepts is found starting from both A and C , and overlap of connected concepts indicates an explanation for the possible association (Weeber et al. 2001).

However, the majority of LBD systems make simplifying approaches about how concepts, and the connections between them, are represented in documents (e.g., Swanson, 1986 or Yetisgen-Yildiz and Pratt 2009). For example, it is commonly assumed that each concept corresponds to a string of text and that concepts which co-occur (e.g., in the same sentence) are connected. These assumptions make it difficult for LBD systems to accurately interpret the information contained in documents. We discuss these sources of difficulty below. Significant advances in the processing of medical and biomedical documents have been made over the last decade but these have not been widely exploited in LBD systems. We suggest ways in which they could be applied to improve LBD by interpreting the information contained in documents more accurately.

Limitations and Possible Solutions

1. Lexical ambiguity: LBD systems often assume each con-

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

cept corresponds to a string. However, this fails to account for the fact that language contains synonymous terms that refer to the same concept (such as “myocardial infarction” and “heart attack”) and this can lead to connections between concepts being missed since the LBD system does not recognise that these are the same (e.g., Tsuruoka, Tsujii, and Ananiadou 2008). In addition, this assumption also fails to take account of the fact that many words and phrases are polysemous, for example, the term “cold” has several possible meanings in biomedical documents, including “virus” and “low temperature”. Consequently, LBD systems may identify connections that do not in fact exist by conflating together different meanings of a term.

Many LBD systems have ignored lexical ambiguity completely, e.g. (Cohn, Schvaneveldt, and Widdows 2010; Gordon and Dumais 1998; Swanson 1986; Tsuruoka, Tsujii, and Ananiadou 2008; Yetisgen-Yildiz and Pratt 2009). Weeber et al. 2001 could not replicate known results unless ambiguity was resolved. A potential solution to this would be to carry out word sense disambiguation (WSD) to identify the meaning of each term in a document before carrying out LBD. This would generate a single interpretation for each term in the document which would remove ambiguity and allow synonymous terms to be identified.

2. **Identifying relations:** Techniques for identifying relationships between concepts employed by LBD have often been limited to simple approaches such as direct co-occurrence (Swanson 1986; 1988; Tsuruoka, Tsujii, and Ananiadou 2008; Yetisgen-Yildiz and Pratt 2009) and higher order associations (Cohn, Schvaneveldt, and Widdows 2010; Gordon and Dumais 1998). While this simple approach is straightforward to implement, it suffers from several limitations:

- (a) there is no explanation of *how* the concepts are related (if in fact at all), and
- (b) phenomena such as negation (e.g., *watching the news does not cause measles*) will not be interpreted.

Applying Information Extraction (IE) before LBD has the potential to avoid these problems (Hoffmann, Zhang, and Weld 2010). IE would carry out a more detailed analysis of the connections between the concepts in the text which could provide a more accurate account of which concepts

are connected and what the relation between them is. Recent work has explored how information about relations can be integrated into LBD systems (Cohen et al. 2011).

3. **Interpreting output:** The output of LBD systems often consist of a list of pairs of concepts that may be connected but no attempt is made to explain what the connection between them is. However, users of LBD systems are likely to be interested in the way in which concepts are connected and may be looking for concepts that are connected in a particular way (e.g., a potential cure).

Techniques from data mining could be applied to provide interpretations for the user (Etzioni et al. 2011; Culotta, McCallum, and Betz 2006), possibly by making further use of the information about connections between concepts produced by LBD. Relationship extraction might be used to explain the relationships between A and B and between B and C concepts. To explain an LBD A and C pair, learning techniques may be able to be trained on a non-LBD corpus of A-B-C relations where the relationships between A and B, B and C, and A and C are all known. The technique would then be able to study LBD candidate discoveries and present A and C relationships given certain A and B and certain B and C relationships.

Conclusion

Interpreting the meaning of text is a challenging task which has not yet been solved. However, significant progress in the interpretation of biomedical documents has been made by the Natural Language Processing community over the last decade. Many LBD systems have used relatively simple techniques for interpreting documents which do not make use of the latest technologies. We highlighted some of the ways in which the limits LBD systems and suggest Natural Language Processing technologies that could be applied to improve LBD.

Acknowledgements

Judita Preiss was supported by the EPSRC grant Language Processing for Literature Based Discovery in Medicine, and M. Heidi McClure would like to thank her employer, Intelligent Software Solutions, for encouraging and supporting her research in the areas of LBD and natural language processing.

References

- Cohen, T.; Widdows, D.; Schvaneveldt, R.; and Rindflesch, T. 2011. Finding schizophrenia's prozac: Emergent relational similarity in predication space. In *Proceedings of the Fifth International Symposium on Quantum Interactions*, 48–59.
- Cohn, T.; Schvaneveldt, R.; and Widdows, D. 2010. Reflective random indexing and indirect inference. *Journal of Biomedical Informatics* 43:240–256.
- Culotta, A.; McCallum, A.; and Betz, J. 2006. Integrating probabilistic extraction models and data mining to discover

relations and patterns in text. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*, 296–303.

Etzioni, O.; Fader, A.; Christensen, J.; Soderland, S.; and Mausam. 2011. Open information extraction: The second generation. In *IJCAI*, 3–10.

Gordon, M., and Dumais, S. 1998. Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science* 49(8):674–685.

Hoffmann, R.; Zhang, C.; and Weld, D. 2010. Learning 5000 relational extractors. In *Proc. ACL*, 286–295.

Swanson, D. 1986. Fish oil, Reynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine* 30:7–18.

Swanson, D. 1988. Migraine and magnesium – 11 neglected connections. *Perspectives in Biology and Medicine* 31(4):526–557.

Tsuruoka, Y.; Tsujii, J.; and Ananiadou, S. 2008. Facta: a text search engine for finding associated biomedical concepts. *Bioinformatics* 24(21):2559–2560.

Weeber, M.; Vos, R.; Klein, H.; and de Jong-van den Berg, L. T. W. 2001. Using concepts in literature-based discovery: Simulating Swanson's Reynaud – fish oil and migraine – magnesium discoveries. *Journal of the American Society for Information Science and Technology* 52(7):548–557.

Yetisgen-Yildiz, M., and Pratt, W. 2009. A new evaluation methodology for literature-based discovery. *Journal of Biomedical Informatics* 42(4):633–643.