Research Proposal - Topic Detection and Tracking of Scientific Literature and Research Proposals

Amund Tveit

http://amund.tveit.org
amund@tveit.org
Breidablikkstien 3, N-7019 Trondheim, Norway

Keywords: Natural Language Processing, Topic Detection and Tracking

1 Introduction

How to cope with information overload is becoming an increasingly important problem even for scientists. Search engines such as Scholar, CiteSeer, SmealSearch, Google, MSN and Yahoo tries to solve this problem by indexing (variable size) samples of publicly available texts (in various formats) from the web. With the recent introduction of *desktop search* the user's own data so that search can be performed simultaneously locally on and on the web. The vision of most Internet search services is indexing of the *deep web*, i.e. the data residing in databases that requires authentication (e.g. commercial digital libraries or company/institution intranets). So the search engines' coverage and recall of the world's available information is in general increasing, as well as the precision of search for *individual* documents. But how is the situation when a user tries to get an overview of specific field described in *several documents?*

2 Research Problem

When a search is performed a user typically gets a (ranked) list of 10 *usually independent* links to relevant documents with a corresponding *dynamic document summary* with highlighting of the matching query terms. So far there has been little focus on showing *relationships* between result documents from search, e.g. when searching for statistical evaluation methods in natural language processing it is probably of greater interest to get a *summary* of such methods gathered from several documents rather than a single document. There can of course exist a good survey paper about what you're looking for that should be returned, but due to the large amounts of documents available in many fields even a landmark a paper would only contain a relatively *small sample* of what is written about it, and for many (specific) queries there aren't any survey papers written at all. This leads to the overall research question:

So how can questions for textual information that spans several documents be answered?

In order to constrain the question I choose to investigate how it can be answered *in the context of scientific literature and research proposals.*

3 Topic Detection and Tracking

One approach for handling the presented research question is *multidocument summarization* (MDS) [17], but questions that can be answered from several documents are usually not on the *entire document level*, e.g. when answering questions about common parameter settings for a machine learning algorithm the introduction part of a paper is probably not the place to look, it is usually better to look in the approach, method or result part. A more fine-grained method for doing this than MDS is *topic detection and tracking* (TDT) [1]. An example of TDT could be to search for papers about natural language processing (NLP) and then apply topic detection either unsupervised to try to discern fragments (e.g. phrases, sentences, paragraphs or sections) of the documents into separate topics, or supervised (e.g. look only for the topic parsing methods in the selected NLP papers). When the topics have been discovered (unsupervised) or selected (supervised), they can be followed or tracked in the direction of choice (e.g. chronologically to get a automatically generated historic survey of parsing methods in the selected set of NLP papers).

3.1 Briefly about TDT origin

Topic Detection and Tracking was initiated by DARPA for finding and following new events in streams of broadcast news stories (and more recent also mentioned in monitoring of potential terror-related information). Online demos of TDT systems (for news) are:

- Columbia Newsblaster - `http://www1.cs.columbia.edu/nlp/newsblaster/`.
- NewsInEssence - `http://lada.si.umich.edu:8080/clair/nie1/nie.cgi`

3.2 Briefly about relationships to natural language processing and machine learning

Topic detection and tracking involves methods from both machine learning and natural language processing, e.g.

- classification [12]
- clustering [23,17]
- domain/language models [22,19,11,10]
- information/entity/noun-phrase extraction [24]
- link/event detection [7,8,6,3,4]
- question answering/dialogue processing [18,20]

– scalability issues [21]
– semantics [16,14]
– sentence ordering [2]
– spatio-temporal issues [15,13]
– speech recognition [25]
– statistical methods [9]
– text segmentation [5]

Since there has been little known research on topic detection and tracking on textual documents from scientific literature and research proposals but evidently a need (in particular for biomedical sciences), this direction is selected to be pursued further.

4 5 Year Plan

4.1 Overall Goal

The goal of my research is to investigate how topic detection and tracking methods can be efficiently developed and utilized for handling scientific texts. Example of possible cases include biomedical and computer science research literature, and the closely related category *research proposals* (e.g. to the or other research councils). The results should be publishable results as well as working (online) prototypes (in a similar manner as the EventSeer digital library I have previously created). (Other opportunities include topic tracking of clinical texts from primary or secondary health care, there exists interesting clinical datasets and guidelines - e.g. NEL - at Norwegian Center for Patient Record Research, NTNU)

4.2 Year 1

Write a up-to-date survey about topic detection and tracking (for ACM Computing Surveys), as well as create a working and extendable platform up and running. Apply the first PhD student. Work with integration to the locally developed GeneTUC. Start building networks with leading academic partners of topic detection and tracking (e.g. Columbia University, Carnegie Mellon University, University of Michigan or University College Dublin) as well as industrial partners and possibly research proposal (to EU or NFR) with partners. EventSeer can be used as a testcase for the initial platforms, it contains more structured data than the scientific literature and has substantial user traffic[1] that can be used to empirically test scalability and initial functionality.

4.3 Year 2-3

Together with PhD student (and others in the group) develop methods and a working online prototype of a topic detection and tracking system for biomedical

[1] People from more than 80 countries using it

texts (based on existing cooperation with biologists and medical researchers), two possible subareas are of particular interest: 1) literature about proteins/genes and how they work and interact (with Professor Astrid Lægreid), and 2) literature about epidemiological studies and experiments (with Associate Professor Arild Faxvaag).

4.4 Year 4-5

Investigate how the approaches can be generalized into other domains (e.g. computer science and physics), other genres (in particular research proposals), other languages and how to efficiently integrate TDT with search engines (either locally developed, public domain engines such as Nutch or Lucene, or with search engine industry)

4.5 Expected Results

The deliverables of this research is expected to be methods and working online prototypes of topic detection and tracking for scientific literature (in biomedicine) and research proposals (in computer science). These should either be open source or ventured through NTNU TTO, all documented and empirically tested in reputable journals, conferences and workshops.

5 Other directions of research

Other opportunies for related research include:

- investigate statistical properties of anaphora resolution, i.e. check if Fred Karlsson's work gives evidence of a statistical distribution (analog to Zip's law for anaphora resolution). And can this be used to improve performance of parsers (reduce expansion levels of grammars) - With Martin Thorsen Ranang
- investigate how to parallelize methods for dimensionality reduction of feature vectors in document clustering and classification, in particular extend work on Fastmap algorithm for dimensionality reduction (in contrast to using PCA, LSI or SVD). Parallel hardware support for dimensionality reduction using FPGAs might be a viable approach due to better APIs that shortens development time and makes it possible to relatively quickly test approaches
- investigate how inductive methods can be used for single- and multi-document summarization, i.e. when documents have been semantically parsed and represented as 1st order logic/prolog. possible approaches include applying association rules, structural support vector machines (SVM-struct) and inductive logic programming.
- use machine learning for training stemmers (e.g. norwegian language), motivation: the porter stemmer isn't good in all domains.

– empirically investigate methods for both measuring quality of and creating dynamic document summaries in search, this can possibly be tested in medium-scale on the EventSeer service first and later in large-scale with industry. Cooperation with Aleksander Øhrn or Knut Magne Risvik
– investigate how eventseer can be as a source for finding papers (in computer science), and investigate topic specific crawling methods to gather additional papers (and research proposals)
– investigate methods for multi-word information/phrase extraction, with Rune Saetre
– participate in TUC-projects, e.g. investigate if Martin's Q/A system for "store norske leksikon" (great Norwegian Lexicon) can be transferred to other domains such as Wikipedia or Medical Handbooks (e.g. Norwegian "felleskatalogen" with medication information, or the guideline Norwegian Electronic Doctors handbook).
– work with on language processing of textual clinical data (such as patient records), it is challenging in terms of how to deal with variable data quality, how to detect who says what (e.g. what the patient feels as opposed to what the doctor observes), as well as general (semantic) parsing issues. Participate on developing the locally developed TUC system for clinical data will be of interest.

References

1. J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.
2. Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. Inferring Strategies for Sentence Ordering in Multidocument Summarization. *Journal of Artificial Intelligence Research*, 17:35–55, 2002.
3. Francine Chen, Ayman Farahat, and Thorsten Brants. Story link detection and new event detection are asymmetric. In *Proceedings of Human Language Technology Conference (HLT-NAACL 2003), Edmonton; Canada*, June 2003.
4. Francine Chen, Ayman Farahat, and Thorsten Brants. Multiple Similarity measures and Source-Pair Information in Story Link Detection. In *Proceedings of Human Language Technology Conference (HLT-NAACL 2004)*, 2004.
5. Freddy Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics, Seattle, Washington*, pages 26–33. Morgan Kaufmann Publishers Inc., 2000.
6. Ayman Farahat, Francine Chen, and Thorsten Brants. Optimizing Story Link Detection is not Equivalent to Optimizing New Event Detection . In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 232–239, 2003.
7. Elena Filatova and Vasileios Hatzivassiloglou . Event-Based Extractive Summarization. In *Proceedings of the ACL Workshop on Summarization*, 2004.
8. Elena Filatova and Vasileios Hatzivassiloglou. Domain-Independent Detection, Extraction, and Labeling of Atomic Events. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-03)*, September 2003.

9. Radu Florian, H. Hassan, A. Ittycheriah, Hongyan Jing, Nanda Kambhatla, X. Luo, Nicolas Nicolov, Salim Roukos, , and T. Zhang. A Statistical Model for Multilingual Entity Detection and Tracking. In *Proceedings of Human Language Technology Conference (HLT-NAACL 2004)*, 2004.
10. Wessel Kraaij. *Variations on Language Modeling for Information Retrieval*. PhD thesis, University of Twente, June 2004.
11. Wessel Kraaij and Martijn Spitters. Language models for topic tracking. In Bruce Croft and John Lafferty, editors, *Language Models for Information Retrieval*. Kluwer Academic Publishers, 2003.
12. Giridhar Kumaran and James Allan. Text classification and named entities for new event detection. In *Proceedings of SIGIR 2004, the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 297–304. ACM Press, 2004.
13. Juha Makkonen and Helena Ahonen-Myka. Utilizing Temporal Information in Topic Detection and Tracking. In Traugott Koch and Ingeborg Sølvberg, editors, *Research and Advanced Technology for Digital Libraries, 7th European Conference, ECDL 2003, Trondheim, Norway, August 17-22, 2003, Proceedings*, volume 2769 of *Lecture Notes in Computer Science*, pages 393–404. Springer-Verlag, Heidelberg, 2003.
14. Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. Applying Semantic Classes in Event Detection and Tracking. In Rajeev Sangal and S. M. Bendre, editors, *Proceedings of International Conference on Natural Language Processing (ICON 2002). Mumbai, India*, 2002.
15. Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. Topic detection and tracking with spatio-temporal evidence. In Fabrizio Sebastiani, editor, *Proceedings of 25th European Conference on Information Retrieval Research (ECIR 2003). Pisa, Italy*, 2003.
16. Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. Simple Semantics in Topic Detection and Tracking. *Information Retrieval (Kluwer)*, 7(3–4):347–368, 2004.
17. Manuel J. Mana-Lopez, Manuel De Buenaga, and Jose M. Gomez-Hidalgo. Multidocument summarization: An added value to clustering in interactive retrieval. *CM Transactions on Information Systems (TOIS)*, 22(2):215–241, April 2004.
18. Mark T. Maybury. Toward a Question Answering Roadmap. Technical report, The MITRE Corporation, November 2002.
19. Ramesh Nallapati. Semantic Language Models for Topic Detection and Tracking. In *Proceedings of Human Language Technology Conference (HLT-NAACL 2004), Student Research Workshop*, 2003.
20. Andrei Popescu-Belis, Alexander Clark Maria Georgescul, Marianne Starlander, and Sandrine Zufferey. A Thematic Bibliography on Dialogue Processing. Technical Report IM2.MDM-06, ISSCO/TIM/ETI, Universit de Geneve, June 2003.
21. K. Seymore and R. Rosenfeld. Large-scale Topic Detection and Language Model Adaptation. Technical Report CMU-CS-97-152, Computer Science Department, Carnegie Mellon University, 1997 1997.
22. Yoshimi Suzuki, Fumiyo Fukumoto, and Yoshihiro Sekiguchi. Topic Tracking Using Subject Templates. In *Proceedings of the 7th International Conference on Spoken Language Processing*, September 2002.
23. Dolf Trieschnigg and Wessel Kraaij. Hierarchical topic detection in large digital news archives. In *Proceedings of the 5th Dutch Belgian Information Retrieval workshop (DIR)*, 2005.

24. Michael White, Tanya Korelsky, Claire Cardie, Vincent Ng, David Pierce, and Kiri Wagstaff. Multidocument Summarization via Information Extraction. In *Proceedings of Human Language Technology Conference (HLT 2001)*, 2001.
25. Lexing Xie, Lyndon Kennedy, Shih-Fu Chang, Ajay Divakaran, Huifang Sun, and Ching-Yung Lin. Layered Dynamic Mixture Model For Pattern Discovery In Asynchronous Multi-Modal Streams. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2005.