Report on the ACM International Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications (ELECTRA 2005) held at SIGIR 2005

Olga Vechtomova
University of Waterloo, Canada
ovechtom@engmail.uwaterloo.ca

Rosie Jones
Yahoo! Overture Matching Sciences, USA
jonesr@yahoo-inc.com

Gaël Dias
University of Beira Interior, Portugal
ddg@di.ubi.pt

1. Introduction

The workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-World Applications (ELECTRA 2005) was held in Salvador, Brazil on August 19 in conjunction with the 28th ACM Annual International Conference on Research and Development in Information Retrieval (SIGIR 2005). The aim of the workshop was to bring together researchers in NLP and IR to discuss the use of lexical cohesion in text applications, such as document and passage retrieval, question answering, topic segmentation and text summarization. There were a number of related workshops in the past: MEMURA 2004 Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications in association with the 4th International Conference on Language Resources (LREC 2004) [1], Multiword Expressions: Integrating Processing Workshop in association with the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004) [2], and Multiword Expressions: Analysis, Acquisition and Treatment Workshop in association with the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003) [3]. The goal of this workshop was to address a wider range of lexical cohesion phenomena in text, not only multiword units, but also relations between words on the sentence, passage and document level, and how they can be useful for information retrieval applications.

We received 12 submissions, 6 of which were accepted as full papers and 3 as short papers. Each paper underwent blind reviewing process by 3 referees. The program committee was composed of 16 members, and included many prominent researchers from both IR and NLP communities. The workshop had 24 participants from 12 different countries (Brazil, Canada, Cuba, Finland, France, India, the Netherlands, Portugal, Sweden, Turkey, United Kingdom and USA). The participants included both academic and industry researchers.

2. Topics

As stated in the call for papers, the main topics of interest in the workshop were:

- Efficient Techniques and scalable algorithms for lexical cohesion identification
- Lexical associations and lexical relations resources
- Document representation and lexical associations
- Document ranking and lexical associations
- Single-term and phrase information retrieval
- Passage retrieval and lexical cohesion

- Query expansion and lexical associations
- Local and global context analysis
- Ontology-based query expansion
- Question answering and lexical relations
- Web search and lexical cohesion
- Topic segmentation and lexical cohesion
- Text summarization and lexical cohesion
- Evaluation standards and benchmarks
- Qualitative and quantitative evaluations

Submissions covered a wide range of these topics. The authors represented NLP and IR communities, bringing different perspectives on these problems.

3. Program Overview

The workshop had a full-day program, which started with a keynote talk by Bruce Croft, followed by 2 full-paper sessions, a poster and demo session and a short-paper session. Each full- and short-paper session was followed by a panel discussion, where the authors of the presented papers were invited to discuss questions pertinent to the applications of lexical cohesion in information retrieval.

The keynote talk by Bruce Croft, Distinguished Professor and Chair, Department of Computer Science, and Director, Center for Intelligent Information Retrieval, University of Massachusetts, was titled "Phrases and Other Structure in Queries". Prof. Croft gave a comprehensive overview of word relationship techniques in Information Retrieval. He discussed the use of phrase indexing, passage retrieval, word co-occurrence, relevance feedback, language models and other techniques in IR. He noted that while for simple queries the bag-of-words approach works adequately well, for complex queries, whose meaning is formed by relationships between concepts and topics, it is important to capture word relationships.

In "Comparing Query Formulation and Lexical Affinity Replacements in Passage Retrieval" Egidio Terra and Charles Clarke compared query formulation strategies and expansion based on lexical affinities in the context of passage retrieval. Their method to expand the queries using lexical affinities replaces only the missing terms from the original query in candidate passages while scoring them. The replacement term's affinity with the missing term is used to weight the substitution, and the degree of affinity is computed using statistics generated from a terabyte corpus. The passages extracted using this replacement method and a set of passages extracted using different formulation strategies are evaluated using TREC's QA test set.

In "A Study of Document Relevance and Lexical Cohesion between Query Terms" Olga Vechtomova, Murat Karamuftuoglu and Stephen Robertson empirically investigated whether the degree of lexical cohesion between the contexts of query terms' occurrences in a document is related to its relevance to the query. Experiments suggest that significant differences between the lexical cohesion in relevant and non-relevant document sets exist. Document re-ranking experiments suggest that lexical cohesion between query terms, formed by the repetition of their collocates, is a useful factor, contributing to improved performance over a document ranking model based on single term weighting.

Bill Hollingsworth and Simone Teufel addressed the problem of how to directly evaluate the quality of lexical chains, in comparison to a human gold standard in "Human annotation of lexical chains: coverage

and agreement measures". Their paper presents a small user study of human annotation of lexical chains, and a set of measures to measure how much agreement between sets of lexical chains there is. They also reported a small metaevaluation to compare the best of these metrics, a partial overlap measure, to rankings of chains derived by introspection, which shows that their measure agrees reasonably well with intuition. They described their algorithm for chain creation, which varies from previous work in several aspects (for instance the fact that it allows for adjective attribution), and reported its agreement with the human annotators in terms of their new measure.

In "A method to calculate probability and expected document frequency of discontinued word sequences" Antoine Doucet and Helena Ahonen-Myka presented a novel technique for calculating the probability of occurrence of a discontinued sequence of n words, that is, the probability that those words occur, and that they occur in a given order, regardless of which and how many other words may occur between them. Their method relies on the formalization of word occurrences into a Markov chain model. Numerous techniques of probability and linear algebra theory are exploited to offer an algorithm of competitive computational complexity. The technique is further extended to permit the calculation of the expected document frequency of an n-words sequence in an efficient manner.

Gaël Dias and Elsa Alves described an innovative Topic Segmentation system in "Unsupervised Topic Segmentation Based on Word Cooccurrence and Multi-Word Units for Text Summarization". Their system is based on a new informative similarity measure based on word co-occurrences, and they evaluate it on a set of web documents within which Multiword Units have previously been identified.

Leif Grönqvist presented a short paper "An evaluation of Bi- and Trigram Enriched Latent Semantic Vector Models". The main motivation for this work was to find an appropriate way to include multi-word units in a latent-semantic vector model. He presented a Swedish evaluation set based on synonym tests and an evaluation of vector models trained with different corpora and parameter settings. The best results in the evaluation were achieved when both bi- and trigrams were added to the models.

In the short paper "Feature Representation for Effective Action-Item Detection", Paul Bennett and Jaime Carbonell discussed a new task of automated action-item detection, with the purpose of flagging emails that require responses, and highlighting specific passages indicating the request for action.

The short paper "Predicting Extraction Performance using Context Language Models" by Eugene Agichtein and Silviu Cucerzan presents a general language modeling method for quantifying the difficulty of information extraction tasks. It demonstrates the viability of their approach by predicting extraction performance of two real world tasks, Named Entity Recognition and Relation Extraction.

The poster and demo session provided an opportunity for more informal discussions. Cyril Goutte presented a poster on behalf of Ágnes Sándor "A framework for detecting conceptual concepts in text". Ricardo Campos presented a poster, co-authored with Gaël Dias, "Automatic Hierarchical Clustering of Web Pages". Leif Grönqvist presented a poster "Evaluating Latent Semantic Vector Models with Synonym Tests and Document Retrieval". In addition, several authors of the presented papers also showed demos of their systems.

4. Conclusions

The workshop provided a stimulating environment for sharing ideas on the applications of lexical cohesion in Information Retrieval. Although there have been many attempts to go beyond the bag of words approach in IR by using lexical cohesion at various levels (phrase, sentence, passage and

document), we felt that there was a need to bring together IR and NLP researchers to discuss what lessons are learned from the past research, which techniques hold more promise, and where should more effort be directed in the future. We feel that the workshop was successful and fulfilled its intended purpose. At the end of the day each participant gave feedback on the workshop. Everyone expressed positive opinion about the workshop, and several people wanted to see it repeating again next year. Some suggestions included a TREC-style task for the participants to work on prior to the workshop, and extending the poster and demo session. The proceedings of the workshop are available online at http://research.yahoo.com/workshops/electra2005/.

5. References

1. Dias, G., Lopes, J.G.L. & Vintar, S. (Eds.) Proceedings of the LREC 2004 Workshop on Methodologies and Evaluation of Multiword Units in Real-world Applications (MEMURA), May 2004, Lisbon, Portugal, ELRA editions.
2. Tanaka T., Villavicencio A., Bond F. and Korhonen A. (Eds.) Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing, July 2004, Barcelona, Spain, ACL press.
3. Bond F., Korhonen A., McCarthy D. and Villavicencio A. (Eds.) Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, July 2003, Sapporo, Japan, ACL press.