Tutorial Proposal for ICON 2013

Title of the Proposed Tutorial: Lexical Chains: Algorithms and Applications
Speaker: Girish Keshav Palshikar
Duration of the Proposed Tutorial: 2 hours (is flexible)

Abstract:
A document is not a sequence of random sentences. Coherence of a document indicates how well the sentences in it "hang together" and give "unity" to the document. Quantifying the coherence of a given document is an important question. Morris and Hirst (1991) introduced lexical chains as one way for measuring the lexical cohesion in a document. A *lexical chain (LC)* is simply an ordered sequence of semantically related words in the document. A LC usually indicates a specific concept as all words within a single LC are related to each other through lexical relations such as repetition, synonymity, hyponymy and being "siblings" in the hypernym tree. Example:

With winds near 180 mph, and even higher gusts, Hurricane Mitch was a Category 5 monster – the highest, most dangerous rating for a storm. The 350-mile wide hurricane was moving west at 8 mph. The U.S. National Weather Service in Miami said Mitch could weaken somewhat, but would still remain a dangerous hurricane.

In general, a document has multiple LCs and analyzing the occurrence of LCs occurring in a document has found many uses, such as automatic summarization, word sense disambiguation, text segmentation, document hyperlinking and topic tracking. Most LC creation algorithms use a lexical resource such as WordNet; they also use criteria such as chain strength to remove some LCs. In this tutorial, we will cover some prominent algorithms for LC creation: Hirst and St-Onge (1998), Silber and McCoy (2002), Galley and McKeown (2003) and Remus and Biemann (2013). We will also briefly cover how LCs have been used for text summarization and document hyperlinking.

References

[1] G. Hirst, D. St-Onge, Lexical Chains as representation of context for the detection and correction malapropisms, in C. Fellbaum (ed.), WordNet: An Electronic Lexical Database, Language, Speech, and Communication, pp. 305–332, The MIT Press, 1998.
[2] H.G. Silber, K.F. McCoy, Efficiently computed lexical chains as an intermediate representation for automatic text summarization, Computational Linguistics, 28(4), pp. 487–496, 2002.
[3] M. Galley, K. McKeown, Improving word sense disambiguation in lexical chaining, Proc. 18th International Joint Conference on Artificial Intelligence (IJCAI'03), pp. 1486–1488, 2003.
[4] S. Remus, C. Biemann, Three Knowledge-Free Methods for Automatic Lexical Chain Extraction, Proc. NAACL 2013.

Outline of the Tutorial

1. Introduction: Cohesion of a document. Lexical chains as a way to measure lexical cohesion. Examples. Summary of approaches to compute lexical chains. Summary of applications.
2. Some prominent algorithms for computing lexical chains in a document
    a. Hirst and St-Onge (1998)
    b. Silber and McCoy (2002)
    c. Galley and McKeown (2003)
    d. Remus and Biemann (2013)
3. Evaluation of lexical chains
4. Applications of lexical chains

    a. Document summarization
    b. Word sense disambiguation
    c. Other applications
5. Computing lexical chains for documents in Indian languages
6. Summary and conclusions
7. References

Brief Biography

Girish Keshav Palshikar obtained M.Sc. (Physics) from Indian Institute of Technology, Bombay in 1985 and an M.S. (Computer Science and Engineering) from Indian Institute of Technology, Chennai in Jan. 1988. Since 1992, he is working in the Tata Research Development and Design Centre (TRDDC), Pune, India, where he is now a *principal scientist* and leads the Machine Learning R&D Group. He was recently honoured with the title of *TCS Distinguished Scientist*. TRDDC is one of the R&D and Innovation Laboratory of Tata Consultancy Services Limited – a premier software company in India. He has about 80 publications in international journals and conferences. He is also a visiting lecturer at the Computer Science Department of University of Pune. His areas of research include machine learning, data and text mining, artificial intelligence and formal methods.

For a full list of publications, please see https://sites.google.com/site/girishpalshikar

Contact Information:
Girish Keshav Palshikar
Tata Research Develeopment and Design Centre (TRDDC),
54B, Hadapsar Industrial Estate,
Pune 411013, India.
Email: GK.Palshikar@tcs.com
Telephone: 91-20-6608 6333 ext. 6400
Fax: 91-20-6608 6399