Automatic Identification of Significant Topics in Domain-independent Full Text Documents

Contact Information

Judith L. Klavans
Center for Research on Information Access
Columbia University
535 W. 114th Street, MC 1101
New York, NY 10027
Phone: 212-854-7443
Fax: 212-854-9099
klavans@cs.columbia.edu

Nina Wacholder
Center for Research on Information Access
Columbia University
535 W. 114th Street, MC 1101
New York, NY 10027
Phone: 212-939-7119
Fax: 212-666-0140
nina@cs.columbia.edu

WWW Page: http://www.columbia.edu/cria/

Keywords: Natural language processing, information access, information retrieval, document analysis, topic identification, topic detection and tracking, text retrieval

Project Award Information:
Award Number: IRI-97-12069
Title: Automatic Identification of Significant Topics in Domain-Independent Full Text Documents
Duration: Three years
Dates: September 1997 to August 2000

Project Summary

The goal of this project is to develop a suite of techniques for identifying significant topics in edited documents such as newspaper articles. Significant entities and concepts are most often referred to in text with nominal expressions such as nouns phrases (e.g., computer science) and proper names (e.g., Buffalo Bills). However, achieving even shallow understanding of nominal expressions remains a major blocking point for natural language processing (NLP). Related to this is the question of the role of natural language information in information retrieval (IR) and specifically the role of nominals in information access. This research is useful for applications such as automatic indexing tasks, either for digital library applications or for direct user manipulation, second stage information retrieval where a subset of a larger corpus has been determined to be potentially relevant, advanced information extraction where important entities in the document must be identified and linked, summarization, and topic detection and tracking.

We have developed linguistically motivated techniques which can link related nominal expressions, to identify documents or portions of documents that are relevant for a given application or user. To the extent that our techniques are based on linguistically-motivated patterns and not on domain-dependent vocabularies, our patterns apply across to text in any genre or domain. We assume that efficient and effective known statistical techniques will continue to be used for large scale tasks such as analysis of large corpora and retrieving a set of possibly relevant documents from the web. One of the hypotheses of our work is that better coverage and better portability across domains and genres is achievable using creative combinations of statistically and linguistically motivated techniques.

Goals, Objectives, and Targeted Activities

To achieve our goal, we are developing:

1. a domain-general method for identifying a list of candidate significant topics in a document that is as complete as is practical, given that full natural language understanding is not likely to be achieved by NLP systems for the foreseeable future.
2. a suite of functions for identifying the most significant of these candidate topics. Our particular contribution solves some of the more challenging problems associated with nominal referring expressions and their variants.
3. a set of tools for evaulating and analyzing the contribution of linguistic information to statistical information retrieval systems.

Indications of Success

In the first year of the project, we developed a modular system, *LinkIT*. The first module of LinkIT, called SNAP, builds a list of candidate significant topics for each document consisting of the complete list of Simplex Noun Phrases (SNPs) (Wacholder 1998). A SNP is a maximal noun phrase with a common noun as its head, where the NP may include premodifiers such as determiners and possessives but not post-nominal constituents such as prepositions or relativizers. Examples are *asbestos fiber* and *9.8 billion Kent cigarette*s. Simplex NPs can be contrasted with complex NPs such as *9.8 billion Kent cigarettes with the filters* where the head of the NP is followed by a preposition, or *9.8 billion Kent cigarettes sold by the company*, where the head is followed by a participial verb. LinkIT then sorts the Simplex Noun Phrases to link together expressions that refer to the same higher level concept such as *rights* in *reproduction rights* and *literary rights*.

The year we have achieved:

1. Refinement of the SNAP module
   - development of a filtering and ranking metric to identify the most significant of the SNPs, based on frequency of the head of the SNP and on frequency of the modifier
   - evaluation of the ranking metric and comparison with two related techniques
2. Implementation of a new module to identify main verbs in documents (VERSO)
3. Analysis of the role of grammatical categories in a statistical IR system (Wacholder et al. under review)
   - creation of set of resources used for training, testing and evalution, including a tagged 330 MB corpus based on the Text Retrieval Conference (TREC) collections
   - development of the DFI (Distance from Ideal) metric to closely analyze the performance of SMART of different versions of documents on specific queries (Evans et al. under review).
4. Refinement of a visualization tool for displaying marked up documents.

LinkIT is used by the two NSF sponsored stimulate projects for the computation of paragraph similarity (Eskin et al. under review) and for the analysis of image captions for building an ontology of multimedia objects (Chang et al. under review). It is also input to a distributed data base project to find background information on the web about significant topics in a document.

If our hypotheses are correct, we will have developed new ways to identify significant topics; we will also have achieved a better understanding of the relative strengths and weaknesses of statistically-based and rule-based natural language processing systems.

Project Impact

Human Resources

- David K. Evans, Ph.D. student, has developed the LinkIT software and the Distance From Ideal (DFI) metric used in evaluation of LinkIT.
- Sonja Allin, B.A. student, major in computer science, has completed a project to improve and refine our tool for document visualization.

Your department/institution infrastructure

This project is part of the Columbia University Digital Library program under the Center for Research on Information Access (CRIA). CRIA sponsors projects to develop tools and techniques which are essential for effective user-oriented text analysis and retrieval, and are also useful for publishing, library, electronic commerce, and other information management tasks in distributed networked environments.

Industry -- collaborations, transfer of technology, patents

We have developed a plan with the Columbia University Libraries and with electronic publishing division of Columbia University Press (CUP) to use *LinkIT* to build an intelligent indexing tool. We are also in contact with the Columbia Law School for potential use in the Columbia Law Review.

What activities have been enabled/spawned because of the accomplishments made possible by your award?

- Eleazar Eskin, a CS Ph.D. student, used LinkIT output as a feature in a machine learning approach to paragraph similarity.
- Carl Sable, a CS Ph.D. student, used LinkIT to help in caption analysis for image classification.
- Stefan Negrila, a Columbia undergraduate, used LinkIT output to cluster documents that discussed the same event. (Supervised by Luis Gravano)

Project References

1. Chang, McKeown, Kathy and all the authors, this is the paper submitted to AAAI on Stim 1 (under review) "Paper title here"
2. Eskin, Eleazar, Judith Klavans, and Vasileios Hatzivassiloglou (under review). "Paper title here".
3. Evans, David Kirk, Judith L. Klavans, and Nina Wacholder (under review). "The Impact of Document Collection Characteristics on Information Access in Digital Libraries",
4. Klavans, Judith L. (1998) "Databases in Digital Libraries: Where Computer Science and Information Management Meet.", ACM-PODS Invited Tutorial, available at "Could you put the url here! ".
5. McKeown, Kathy and all the authors, this is the paper submitted to AAAI on Stim 2 (under review)

"Paper title here"
6. Wacholder, Nina, Judith L. Klavans and David Kirk Evans (under review) "The role of grammatical categories in a statistical information retrieval system"
7. Wacholder, Nina (1998). "Simplex NPS Sorted by Head: a Method for Identifying Significant Topics within a Document," *Proceddings of the COLING-ACL Workshop on the Computational Treatment of Nominals*, Montreal, Canada, August 16, 1998.
8. Wacholder, Nina, Yael Ravin and Misook Choi (1997) "Disambiguation of proper names in text," *Proceedings of the ANLP*, ACL, Washington, DC.

Area Background

The NSF is actively involved in funding research to enable increased universal access to the fast-growing body of electronic text. Our research directly addresses the need to find information more easily and more reliably. We are devloping a range of innovative methods to improve current methodologies for information retrieval, indexing, extraction, and summarization.

Area References

Losee, Robert. 199X Dave, another ref to fill out!
Klavans, Judith L. and Philip Resnik, editors (1996*). The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. MIT Press, Cambridge, Mass.
