Linking Content in Unstructured Sources

Marie-Francine Moens
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
B-3001 Heverlee, Belgium
sien.moens@cs.kuleuven.be

ABSTRACT

This tutorial focuses on the task of automated information linking in text and multimedia sources. In any task where information is fused from different sources, this linking is a necessary step. To solve the problem we borrow methods from computational linguistics, computer vision and data mining. Although the main focus is on finding equivalence relations in the sources, the tutorial opens views on the recognition of other relation types.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: linguistic processing; I.2.7 [**Natural Language Processing**]: text analysis, machine translation; I.2.10 [**Vision and Scene Understanding**]: video analysis

General Terms

Algorithms, Experimentation

Keywords

Content alignment and linking

1. INTRODUCTION

We witness a growing interest and capabilities of automatic content recognition (often referred to as information extraction) in various unstructured media sources that identify entities (e.g. persons, locations and products) and their semantic attributes (e.g., opinions expressed towards persons or products, relations between entities). These extraction techniques are most advanced for text sources, but they are also researched for other media, for instance, for recognizing persons and objects in images or video. The extracted information enriches and adds semantic meaning to documents and queries in a search setting. An important challenge is to automatically link equivalent and complementary content allowing for an intelligent fusion of information and reasoning across documents, Web pages and other information sources. We succeed quite successfully in assigning subject categories to content in a supervised learning setting, when sufficient training examples are available and where the subject categories act as an interlingua. However, automatically linking content on a more detailed level has received lesser attention, but is equally important and poses a number of opportunities and challenges. For instance, we can detect when a name refers to the same person, so we can couple the person's birthdate found on a home page with the identified hobbies of that person found in blogs, and perhaps recognize the person in a picture on another Web page. We might recognize factors in news that describe a business in trouble, while similar information can be found in other wordings for another company in blogs.

The World Wide Web is very diverse covering many different languages, media and disciplines. A first challenge is the development of generic algorithms for linking content across documents, languages and media. Emphasis is on joint processing of the different modalities and on generic algorithms for linking content. Another challenge is to develop technologies where the manual human effort or supervision is minimal. Content linking is a timely topic because in any usage of information we couple data together and make inferences, but today there are a number of techniques for alignment of content developed in the computational linguistics, computer vision and data mining communities, that are worth studying and that have a proven usefulness in heterogeneous settings. The tutorial goes deeper into current approaches of automated linking, including probabilistic methods that maximize the likelihood of aligning recognized content. The results will increase the linked data on the World Wide Web. Linked data are a necessary prerequisite for many applications such as Web mining, question answering search on the Web, search based on link analysis models, summarization of Web data and many others.

The tutorial gives an overview of current information linking techniques for unstructured data such as text and images, and is illustrated with examples in monolingual, cross-lingual and cross-media settings. Among the motivating example applications are paraphrasing, person search, cross-lingual term extraction and event linking, and cross-media entity alignment.

2. GOALS AND OUTCOME

The tutorial's main goal is to give the participants a clear and detailed overview of content linking approaches and tools, and the integration of their results into several applications in the framework of Web search, mining and summarization. A small set of integrated and interactive exercises will sharpen the understanding by the audience. By attending the tutorial, attendants will:

- Acquire an understanding of alignment algorithms for content linking;

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

- Acquire an understanding of the integration of constraints in the alignment algorithms;
- Be able to use the models for different languages and media and be aware of the necessary preprocessing in these modalities;
- Be able to choose a model for content linking that is well-suited for a particular task or application.

3. COURSE CONTENT

The tutorial will consist of the following parts:

1. Motivation: developments in automated linking of content in computational linguistics, computer vision and data mining; potential for information access, mining and summarization; introduction to the applications;
2. Probability theory, notations, and basic concepts including topic models, statistical alignment models, approximate inference and the expectation maximization algorithm;
3. Monolingual linking of content:
   (a) Within document noun phrase coreferent resolution: constrained clustering, experiences from data mining and natural language processing;
   (b) Cross document noun phrase coreferent resolution: problems of polysemy and synonymy; illustrated with Web people search and event completion;
   (c) Paraphrasing: unsupervised learning of paraphrases, latent words language modeling, illustrated with multidocument summarization;
4. Cross-lingual linking of content:
   (a) Alignment in parallel texts: association metrics, IBM models, Hidden Markov Models, asymmetric and symmetric models, generative models and discriminative models; illustration with cross-lingual term extraction;
   (b) Alignment in comparable texts; illustrated with cross-lingual topic detection and clustering;
5. Cross-media linking of content:
   (a) Alignment with constrained expectation maximization and deterministic annealing; illustrated with alignment of names and faces in Web pages;
6. Conclusions and future perspectives.

The tutorial lasts for 3 hours. Parts 1 and 2 of the tutorial take up each about 20 minutes. Parts 3 and 4 will last each about 45 minutes.There will be a break between parts 3 and 4 of about 10 minutes. Part 5 will take up 30 minutes and another 10 minutes is reserved for part 6. The format of the tutorial consists of lectures, open discussions and exercises in which the tutorial participants will discuss and apply the lessons learned.

4. COURSE MATERIAL

Handouts of slides, and a detailed bibliography will be available for the participants of the tutorial. If needed, for instance, based on discussions on site, additional information will be made available on the World Wide Web.

5. TUTORIAL AUDIENCE

The tutorial is aimed at students, teachers, and academic and company researchers who want to gain an understanding of current information linking technologies that automatically enrich text and multimedia content, and of several illustrating applications. As such, the tutoral might also be relevant for developers of Semantic Web applications. An understanding of common data mining techniques is recommended.

6. BIOGRAPHY

Marie-Francine Moens is a tenured associate professor at the Department of Computer Science of the Katholieke Universiteit Leuven, Belgium. She holds a Ph.D. degree in Computer Science (1999) from this university. She currently leads the research team *Language Intelligence and Information Retrieval* composed of 1 postdoctoral researcher and 10 doctoral students, and is currently coordinator of or partner in numerous international and European research projects (FP6, FP7, ITEA2) in the fields of information retrieval, natural language processing and text mining. Her main interests are in the domain of automated content retrieval from texts with a strong emphasis on probabilistic content models obtained through machine learning techniques. Since 2001 she teaches the course *Text Based Information Retrieval* and since 2009 she partly teaches the courses *Natural Language Processing* and *Current Trends in Databases* at K.U.Leuven. In 2008 she lectured the course *Text Mining, Information and Fact Extraction* at *RuSSIR'2008: the 2nd Russian Summer School in Information Retrieval.* She is author of 170 international publications among which are two monographs published by Springer, including a book on information extraction. She is (co-)editor of 12 books or proceedings, (co)-author of 25 international journal articles and 20 book chapters. She is involved in the organization or program committee of major conferences on computational linguistics and information retrieval (ACL, COLING, EACL, SIGIR, ECIR, CIKM). She is the (co-)organizer of 2 editions of the DIR - *Dutch-Belgian Information Retrieval Workshop*, 3 editions of the KRAQ - *Knowledge and Reasoning for Answering Questions* conferences (respectively at IJCAI 2005, COLING 2008 and ACL 2009), the *ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics* (KDD 2009) and the *Cross-media Information Access and Mining* workshop (IJCAI-AAAI 2009). She is appointed as chair-elect of the European Chapter of the Association for Computational Linguistics (2009-2010).