Information Extraction and Linking in a Retrieval Context

Marie-Francine Moens[1] and Djoerd Hiemstra[2]

[1] Katholieke Universiteit Leuven,
Department of Computer Science,
Celestijnenlaan 200A,
B-3001 Heverlee, Belgium
sien.moens@cs.kuleuven.be
[2] University of Twente
Department of Computer Science,
PO Box 217, 7500 AE,
Enschede, The Netherlands
d.hiemstra@cs.utwente.nl

1 Introduction

We witness a growing interest and capabilities of automatic content recognition (often referred to as information extraction) in various 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 document and queries (the latter e.g., in a relevance feedback setting). In addition, content recognition techniques trigger automated linking of information across documents and even across media. This situation poses a number of opportunities and challenges for retrieval and ranking models. For instance, instead of returning full documents, information extraction provides the means to return very focused results in the form of entities such as persons and locations. Another challenge is to integrate content recognition and content retrieval as much as possible, for instance by using the probabilistic output from the information extraction tools in the retrieval phase. These approaches are important steps towards semantic search, i.e., retrieval approaches that truly use the semantics of the data.

We propose a half day tutorial which gives an overview of current information extraction techniques for text, including among others entity recognition and entity relation recognition. Examples of content recognition in other media are given. The tutorial goes deeper into current approaches of automated linking, including probabilistic methods that maximize the likelihood of aligning recognized content. As a result, documents can be modeled as mixtures of content, incorporating certain dependencies, and document collections can be represented as a web of information. An important part of the tutorial focuses

M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 810–813, 2009.

on retrieval models and ranking functions that use results of the information extraction. We explain the use of probabilistic models, more specifically relevance language models for entitity retrieval, graph models and probabilistic random walk models for entity retrieval, and extensions of models to handle noisy entity recognition or noisy concept recognition. The tutorial includes several motivating examples and applications among which are expert search using output from named entity tagging, connecting names to faces in videos for person search using output from named entity tagging and face detection, video search using output from concept detectors, and spoken document retrieval using speech lattices and posterior probabilities of recognized words. The examples will be combined in a larger case study: *Retrieval of news broadcast video.*

2 Goals and Outcome

The tutorial's main goal is to give the participants a clear and detailed overview of content modeling approaches and tools, and the integration of their results into ranking functions. 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 current information extraction, topic modeling and entity linking techniques;
- Acquire an understanding of ranking models in information retrieval;
- Be able to integrate the (probabilistic) content models into the ranking models;
- Be able to choose a model for retrieval that is well-suited for a particular task and to integrate the necessary content models.

3 Course Content

The tutorial will consists of the following parts:

1. Motivation: developments in content recognition (computational linguistics, computer vision, audio processing), possibilities of automatically linking equivalent content, potential for information access and retrieval, introduction to the applications (by Moens);
2. Probability theory, notations, and basic concepts including language models and the Robertson/Sparck-Jones probabilistic model (by Hiemstra);
3. Emerging information extraction and linking techniques that semantically enrich the data sources (mainly text): named entity recognition, cross-document co-reference resolution, entity linking by expectation maximization, while not neglecting the natural language characteristics (e.g. obtained by partof- speech tagging and shallow parsing) used and the content representations (by Moens);
4. Ranking models: extensions of ranking models for entity search and noisy concept recognition, amongst others: relevance models, random walk models, and extended probabilistic models; (by Hiemstra)

5. Case study: Retrieval of news broadcast video. Recognizing names of persons and locations, recognizing concepts such as faces and aligning person names to faces. Using noisy annotations to search for videos (by Moens and Hiemstra).

Part 1 and 2 are treated together as one course part. These parts take about 20 minutes each to get all participants on the same level of basic knowledge of content recognition and retrieval models. Combined approaches will be presented in Part 3 and Part 4. Each of the four parts will take about 45 minutes (depending on the organizations tutorial schedule), with breaks in between. The case study will consist of 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 extraction technologies that automatically enrich text and multimedia data with semantics, the integration of the extraction technologies into ranking models for information retrieval, and of several illustrating retrieval applications. As such, the tutoral might also be relevant for developers of Semantic Web applications.

6 Biographies

Marie-Francine Moens is 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 a research team of 2 postdoctoral fellows and 8 doctoral students, and is currently coordinator of or partner in 7 European research projects in the fields of information retrieval 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 *RuSSIR2008*: *the 2nd Russian Summer School in Information Retrieval*. She has (co-)authored more than 130 research papers in the field of IR and text analysis, and is author of two monographs published in the *Springer International Series*

on Information Retrieval. She is the (co-)organizer of 2 editions of the *DIR - Dutch-Belgian Information Retrieval Workshop* (2002 and 2007), one of which was organized together with Djoerd Hiemstra, 3 editions of the *KRAQ - Knowledge and Reasoning for Answering Questions* conferences (respectively at IJCAI 2005, COLING 2008 and ACL 2009), and the *Cross-media Information Access and Mining workshop* (IJCAI-AAAI 2009). She was recently appointed as chair-elect of the *European Chapter of the Association for Computational Linguistics* (2009-2010).

Djoerd Hiemstra is assistant professor at the Department of Computer Science of the University of Twente in the Netherlands. He contributed to over 100 research papers in the field of IR, covering topics such as language models, structured information retrieval, and multimedia retrieval. Djoerd gave lectures on *Formal models of IR* at two editions of the European Summer School on Information Retrieval (ESSIR). He is focus director *data management, storage and retrieval* of the *Dutch Research School of Information and Knowledge Systems* (SIKS), an interuniversity research school that comprises 12 research groups in which currently nearly 400 researchers are active, including over 190 Ph.D. students. Djoerd is involved in several advanced SIKS courses for Dutch Ph.D. students. Djoerd was involved in the local organization of SIGIR 2007 in Amsterdam, and in the organization of several workshops including three editions of the Dutch- Belgian Information Retrieval Workshop series, one of which organized together with Marie-Francine Moens.