

# Preface

## 1 Introduction

Information extraction (IE) and text summarization (TS) are key technologies aiming at extracting relevant information from texts and presenting the information to the user in condensed form. The on-going information explosion makes IE and TS particularly critical for successful functioning within the information society. These technologies, however, face new challenges with the adoption of the Web 2.0 paradigm (e.g. blogs, wikis) because of their inherent multi-source nature. These technologies have to deal no longer with isolated texts or single narratives, but with large-scale repositories, or sources – possibly in several languages – containing a multiplicity of views, opinions, or commentaries on particular topics, entities or events. There is thus a need to adapt and/or develop new techniques to deal with these new phenomena.

Recognising similar information across different sources and/or in different languages is of paramount importance in this multi-source, multi-lingual context. In information extraction, merging information from multiple sources can lead to increased accuracy relative to extraction from a single source. In text summarization, similar facts found across sources can inform sentence scoring algorithms. In question answering, the distribution of answers in similar contexts can inform answer ranking components.

Often, it is not the similarity of information that matters, but its complementary nature. In a multi-lingual context, information extraction and text summarization can provide solutions for cross-lingual access: key pieces of information can be extracted from different texts in one or many languages, merged, and then conveyed in many natural languages in concise form. Applications need to be able to cope with the idiosyncratic nature of the new Web 2.0 media: mixed input, new jargon, ungrammatical and mixed-language input, emotional discourse, etc. In this context, synthesizing or inferring opinions from multiple sources is a new and exciting challenge for NLP. On another level, profiling of individuals who engage in the new social Web, and identifying whether a particular opinion is appropriate/relevant in a given context are important topics to be addressed.

It is therefore important that the research community address the following issues:

- What methods are appropriate to detect similar/complementary/contradictory information? Are hand-crafted rules and knowledge-rich approaches suitable?
- What methods are available to tackle cross-document and cross-lingual entity and event co-reference?
- What machine learning approaches are most appropriate for this task—supervised, unsupervised, semi-supervised? What type of corpora are required for training and testing?
- What techniques are appropriate to synthesize condensed synopses of the extracted information? What generation techniques are useful here? What kind of techniques can be used to cross domains and languages?
- What techniques can improve opinion mining and sentiment analysis through multi-document analysis? How do information extraction and opinion mining connect?
- What tools exist for supporting multi-lingual/multi-source access to information? What solutions exist beyond full document translation to produce cross-lingual summaries?

This volume contains a series of recent papers covering most of the above challenges. Some of them also bridge the gap between IE and TS and show that these are complementary technologies that can be valuably integrated in real-world natural language applications.

## 2 Content of this volume

Part I of the volume contains two background chapters describing the state of the art in Text Summarization (Saggion and Poibeau) and Information Extraction (Piskorski and Yangarber). These are intended to provide a broad overview of the field for the reader, and to define the context for the subsequent technical chapters.

Part II contains four chapters on named entity analysis in a multilingual context. Named entity recognition plays a prominent role both for IE and TS. Named entities carry major information that can help determine what is the text about. It is a major component of any IE system, which consists in large part in identifying relations between named entities. Lastly, named entity also plays a role in determining what are the most important sentences of a text, which is obviously crucial for TS. This Part of the volume provides different studies on named entity recognition, addressing issues such as variation across languages (Mani et al.) or in one language (Driscoll). The last two chapters address the complex problem of relating various pieces of information to one referring entity despite language variation (Rao et al.) and co-reference resolution across documents (Saggion).

The first chapter is “Learning to Match Names Across Languages” by Inderjeet Mani, Alex Yeh and Sherri Condon. The authors report on research on matching

names in different scripts across languages. They explore two trainable approaches based on comparing pronunciations. The first, a cross-lingual approach, uses an automatic name-matching program that exploits rules based on phonological comparisons of the two languages carried out by humans. The second, monolingual approach, relies only on automatic comparison of the phonological representations of each pair. Alignments produced by each approach are fed to a machine learning algorithm. Results show that the monolingual approach results in machine-learning based comparison of person-names in English and Chinese at an accuracy of over 97.0 F-measure.

The following chapter is “Computational Methods for Name Normalization Using Hypocoristic Personal Name Variants” by Patricia Driscoll. A growing body of research addresses name normalization as part of co-reference and entity resolution systems, but the problem of hypocoristics has not been systematically addressed as a component to such systems. In many languages, these name variants are governed by morphological and morpho-phonological constraints, providing a dataset rich in features which may be used to train and run matching systems. This chapter gives a full treatment to the phenomenon of hypocoristics and presents a supervised learning method that takes advantage of their properties to untangle the relationships between hypocoristics and corresponding full form names.

The following chapter, “Entity Linking: Finding Extracted Entities in a Knowledge Base”, by Delip Rao, Paul McNamee and Mark Dredze, deals with named entity disambiguation as well as connecting various pieces of information from different texts to the same real-world object. Entity linking is a new task that has recently drawn much attention in NLP research. Entity Linking, also referred to as record linkage or entity resolution, involves aligning a textual mention of a named-entity to an appropriate entry in a knowledge base, which may or may not already contain the entity. This has manifold applications, ranging from linking patient health records to maintaining personal credit files, prevention of identity crimes, and supporting law enforcement. The authors discuss key challenges of this task, and present a high-performing system that links entities using max-margin ranking. The chapter also summarizes recent work in this area and describes several open research problems.

Once named entities have been analyzed, they can be used to identify related documents, which involve co-reference resolution. This is the topic addressed by the last chapter of Part II: “A Study of the Effect of Document Representations in Clustering-based Cross-document Co-Reference Resolution” by Horacio Saggion. Finding information about people on huge text collections or on-line repositories on the Web is a common activity. The author describes experiments aiming at identifying the contribution of semantic information (e.g. named entities) and summarization (e.g. sentence extracts) in a cross-document co-reference resolution algorithm. Its cross-document co-reference system is a clustering-based algorithm which groups documents referring to the same entity. Clustering uses vector representations created by summarization and semantic tagging analysis components. The author investigates different configurations achieving state of the art performance demonstrating the potential of the applied techniques. He shows

that selection of the type of summary and the type of term to be used for vector representation is important to achieve good performance.

Part III contains three chapters on Information Extraction. The chapters address various aspects of IE, both practical and theoretical. IE typically refers to filling a predefined template with information extracted from text related to a specific domain. The chapters address IE in less favorable environments: for example, when no domain or no template have been defined, or when the system faces a large number of different domains, or different sources (Neumann and Schmeier). In this context, it can be useful to predict the utility of the extracted information (Huttunen et al.). The output of IE can also serve as a basis for the generation of summaries (Ji et al.).

The first chapter is “Interactive Topic Graph Extraction and Exploration of Web Content” by Günter Neumann and Sven Schmeier. The authors consider IE when no domain has been *a priori* defined. It is then necessary to let the user dynamically explore the corpus and define templates on the fly. In their framework, the initial information request (in the form of a query topic description) is issued by a user online to the system using a search engine. A topic graph is then constructed using collocations identified in snippets returned by the search engine. This graph allows the user to dynamically explore the results, refine his query and identify additional relevant knowledge using the topic graph. The authors conclude their chapter with a user-oriented evaluation, which shows that the approach is especially helpful for finding new interesting information on topics about which the user has only a vague idea or no idea at all.

The following chapter is “Prediction of Utility in Event Extraction”, by Silja Huttunen, Arto Vihavainen, and Roman Yangarber. The goal of the chapter is to estimate the relevance of the results of an Information Extraction system to the end-user’s needs. Traditional criteria for evaluating the performance of IE focus on correctness of the extracted information, e.g., in terms of recall, precision and F-measure. Here, the authors introduce subjective criteria for evaluating the quality of the extracted information: utility of results to the end-user. They use methods from text mining and linguistic analysis to identify features that are good predictors of the relevance of an event or a document to a user. They report on experiments in two real-world news domains: business activities and epidemics of infectious disease.

In “Open-Domain Multi-document Summarization via Information Extraction: Challenges and Prospects”, Heng Ji, Benoit Favre, Wen-Pin Lin, Dan Gillick, Dilek Hakkani-Tur and Ralph Grishman propose ideas to bridge the gap between IE and TS. The authors observe that IE and TS share the same goal of extracting and presenting relevant information in a document. They show that while IE was a primary element of early abstractive summarization systems, it has been left out in more recent extractive systems. However, extracting facts, recognizing entities and events should provide useful information to those systems and help resolve semantic ambiguities that they cannot tackle. The chapter explores novel ways of taking advantage of cross-document IE for multi-document summarization. The authors propose several approaches to IE-based summarization and analyze

their strengths and weaknesses. One of them, re-ranking the output of a high performing summarization system with IE-informed metrics, leads to improvements in manually-evaluated content quality and readability.

The final part, Part IV of this volume, concerns multi-document summarization. There is a clear link between this and the previous parts, since recognition of named entities and identification of key information in text are among the main components of any automatic summarization system. The five chapters of Part IV address recent trends in automatic summarization, such as production of update summaries, containing only new information after a first set of documents has already been summarized (Bossard) and production of summaries in a highly multilingual environment (Kabadjov et al.). An important issue in this context is coherent ordering of the information, especially when information is extracted from multiple sources (Bollegala et al.). The two last contributions consider TS techniques in multimedia environments, namely in relation to speech (Ribeiro and Martins de Matos) and images (Aker et al.).

The first chapter in Part IV is “Generating Update Summaries: Using an Unsupervised Clustering Algorithm to Cluster Sentences” by Aurélien Bossard. The author presents an original approach based on clustering techniques: sentence clustering makes it possible to group sentences conveying the same event or the same idea. In the first step, sentences of an initial set of documents are clustered based on their content. Sentences contained in new documents are then projected onto the result of the clustering of the initial set of documents, making it possible to distinguish new information (merged in existing clusters) from already known information (forming new clusters). The system is evaluated on the TAC 2009 “Update task” and shows promising results.

In “Multilingual Statistical News Summarisation”, Mijail Kabadjov, Josef Steinberger and Ralf Steinberger present a generic approach for summarizing multilingual news clusters, such as those produced by the Europe Media Monitor (EMM) system. The authors use robust statistical techniques to summarize news from different sources and languages. A multilingual entity disambiguation system is used to build the source representation. The authors show that their system obtains good performances on the TAC datasets. Lastly, the authors have run a small-scale evaluation on languages other than English, providing interesting evidence that contradicts the pervasive assumption “if it works for English, it works for any language.”

The following chapter, “Coherent Ordering of Information Extracted from Multiple Sources” by Danushka Bollegala, Naoaki Okazaki and Mitsuru Ishizuka, in creating a summary of information extracted from multiple sources, deals with the problem of deciding on the order in which information must be presented in the summary. Incorrect ordering can lead to misunderstandings. In this chapter, the authors discuss the challenges involved when ordering information selected from multiple sources and present several approaches to overcome those challenges. They also introduce several semi-automatic evaluation measures to empirically evaluate an ordering of sentences created by an algorithm.

The two last chapters establish connection between TS and other media, speech and images. Concerning the former, Ricardo Ribeiro and David Martins de Matos present a chapter entitled “Improving Speech-to-Text Summarization by Using Additional Information Sources”. Speech-to-text summarization systems usually take as input the output of an automatic speech recognition (ASR) system that is affected by speech recognition errors, disfluencies, or difficulties in identification of sentence boundaries. The authors propose the inclusion of related, solid background information to cope with the difficulties of summarizing spoken language and the use of multi-document summarization techniques in single document speech-to-text summarization. They explore the possibilities offered by phonetic information to select the background information and conduct a perceptual evaluation to assess the relevance of the inclusion of that information. Results show that summaries generated using this approach are better than those produced by a state-of-the-art latent semantic analysis (LSA) summarization method and suggest that humans prefer summaries restricted to the information conveyed in the input source.

The last chapter, “Towards automatic image description generation using multi-document summarization techniques” by Ahmet Aker, Laura Plaza, Elena Lloret, and Robert Gaizauskas, reports an initial study of the viability of multi-document summarization techniques for automatic captioning of geo-referenced images. The automatic captioning procedure requires summarizing multiple web documents that contain information related to the images’ location. The authors use different state-of-the-art summarization systems to generate generic and query-based multi-document summaries and evaluate them using ROUGE metrics relative to human-generated summaries. Results show that query-based summaries perform better than generic ones and are thus more appropriate for the task of image captioning, or generation of short descriptions related to the location/place captured in the image.

Montrouge, France  
Barcelona, Spain  
Warszawa, Poland  
Helsinki, Finland

Thierry Poibeau  
Horacio Saggion  
Jakub Piskorski  
Roman Yangarber

Multi-source, Multilingual Information Extraction and  
Summarization

Poibeau, T.; Saggion, H.; Piskorski, J.; Yangarber, R.  
(Eds.)

2013, XX, 324 p., Hardcover

ISBN: 978-3-642-28568-4