Natural Language Processing and the Web

Researchers in a wide range of natural language processing (NLP) sub-areas are increasingly turning to the Web as a data source. Several groups have explored the Web’s potential for building lexical resources or for machine translation, by either creating bilingual corpora or using the Web to filter out or postedit translation candidates. Other research has discovered semantic relations by querying the Web for lexico-syntactic patterns indicative of hyponymy, entailment, similarity, antonymy, or enablement. Studies have investigated the Web’s usefulness for word sense disambiguation, question answering, language modeling, and summarization.

This special issue focuses on applications that innovatively use Web-scale document collections to create useful resources or applications that ultimately let end users navigate the Web more easily. The call for papers attracted about 30 submissions; we’re pleased to present four that cover a variety of novel applications.

In “Mining the Web to Create Specialized Glossaries,” Paola Velardi, Roberto Navigli, and Pierluigi D’Amadio (Univ. of Rome “La Sapienza”) address automatic compilation of domain-specific glossaries from the Web. Terms have different meanings depending on the domain. For example, in the financial domain, “security” refers to investment instruments, whereas in the information technology domain, it refers to the protection of a computer system’s integrity. The authors’ system includes several components, including modules for term extraction, glossary creation based on Web-mining techniques, and glossary filtering and validation.

In “Learning to Tag and Tagging to Learn: A Case Study on Wikipedia,” Peter Mika, Massimiliano Ciaramita, Hugo Zaragoza, and Jordi Atserias (Yahoo! Research, Barcelona) examine named entity recognition (NER), a method for automatically assigning semantic categories to named entities as they appear in running text. Traditionally, NER categories include people, places, organizations, artifacts, and quantities, among others. One of the biggest challenges to NER is that texts in different domains and from different sources use different surface patterns to express the same concept. So, porting a system to a completely new text source requires domain adaptation. In this work, the authors perform domain adaptation and use NER to enrich Wikipedia metadata.

The third article addresses a complementary topic. In “Linking Documents to Encyclopedic Knowledge,” Andras Csomai and Rada Mihalcea (Univ. of North Texas) use Wikipedia to perform automated keyword extraction and word sense disambiguation. A typical task to which they apply their method is text “wikification”—automatically identifying words and phrases in a document and linking each to the appropriate Wikipedia entry. Keyword extraction is needed to address ambiguous phrases. One such example is “New York Times,” which should link to the newspaper’s entry, not the city’s entry. On the other hand,
word sense disambiguation is needed to decide which target pages with the same entry name to link to. For example, Wikipedia has more than 20 different entries labeled “Chicago,” whose topics include not only the eponymous city, musical, movie, and rock band but also a poker game, an asteroid, a ship, and a typeface. The authors address the challenge of finding the right target page.

“Weighing Stars: Aggregating Online Product Reviews for Intelligent E-commerce Applications,” by Zhu Zhang (Univ. of Arizona), deals with subjectivity and sentiment analysis. Zhang aggregates multiple reviews of the same product into a single review by using support vector machines and a number of lexical and syntactic features.

As the best practices in system-based NLP require, all four articles assess in detail the performance of the underlying algorithms and systems, using both user-based and automated evaluation.

Acknowledgments

We thank all the reviewers for the time they devoted to this special issue.

References


