Chapter 13

Disambiguation and Filtering Methods in Using Web Knowledge for Coreference Resolution

Olga Uryupina
University of Trento, Italy

Massimo Poesio
University of Trento, Italy & University of Essex, UK

Claudio Giuliano
University of Trento, Italy & Fondazione Bruno Kessler, Italy

Kateryna Tymoshenko
University of Trento, Italy & Fondazione Bruno Kessler, Italy

ABSTRACT

The authors investigate two publicly available Web knowledge bases, Wikipedia and Yago, in an attempt to leverage semantic information and increase the performance level of a state-of-the-art coreference resolution engine. They extract semantic compatibility and aliasing information from Wikipedia and Yago, and incorporate it into a coreference resolution system. The authors show that using such knowledge with no disambiguation and filtering does not bring any improvement over the baseline, mirroring the previous findings (Ponzetto & Poesio, 2009). They propose, therefore, a number of solutions to reduce the amount of noise coming from Web resources: using disambiguation tools for Wikipedia, pruning Yago to eliminate the most generic categories and imposing additional constraints on affected mentions. The evaluation experiments on the ACE-02 corpus show that the knowledge, extracted from Wikipedia and Yago, improves the system’s performance by 2-3 percentage points.

DOI: 10.4018/978-1-61350-447-5.ch013
INTRODUCTION

Coreference resolution is an essential prerequisite for a variety of NLP tasks: information extraction, machine translation, summarization and many others. State-of-the-art coreference resolvers often rely on very shallow features. Although several models of coreference have been proposed recently, advocating more complex machine learning frameworks, knowledge acquisition still remains a bottleneck for state-of-the-art algorithms.

The Semantic Web has made available a large amount of information, which constitute a valuable source of semantics. However, it’s difficult to integrate it with state-of-the-art coreference methods: different Semantic Web resources use different schemata; knowledge bases have irregular coverage; Semantic Web knowledge is encoded in logical form while coreference systems are based on statistical models.

In this paper we investigate two publicly available web knowledge bases, Wikipedia and Yago, in an attempt to leverage aliasing and semantic information and thus increase the performance level of our coreference resolution engine. We show that a naive approach does not bring any improvement over the baseline, mirroring the previous findings. We propose, however, a number of solutions to reduce the amount of noise coming from web resources: using disambiguation tools for Wikipedia, pruning Yago to eliminate the most populated categories and imposing additional constraints on affected mentions. Our evaluation experiments on the ACE-02 corpus show that the knowledge, extracted from Wikipedia and Yago, improves our system’s performance by 2-3 percentage points.

We also perform an error analysis, identifying cases where semantic compatibility information induced from the web leads to spurious coreference links. It suggests that most errors can be eliminated by taking into account mentions’ modifiers: for example, “a European country” and “Canada” cannot be coreferent, contrary to the prediction of our Yago features. We thus plan to investigate the interaction of syntactic and web-based knowledge to further improve our coreference resolution engine.

The rest of the paper is organized as follows. First, we briefly discuss related studies, then we describe our baseline system. In the following two sections we discuss different methods for integrating Wikipedia and Yago into our baseline system. Finally, we discuss our evaluation results and present the conclusions.

RELATED WORK

Coreference is a complex phenomenon and therefore a robust and reliable approach to the problem should address numerous linguistic and common-sense aspects of the task. Previous studies have investigated possibilities for extracting such knowledge from WordNet (Harabagiu, Bunescu, & Maiorano, 2001, Huang, G., W., & A., 2009), Wikipedia (Ponzetto & Strube, 2006) or large text corpora (Haghighi & Klein, 2009, Bean & Riloff, 2004, Garera & Yarowsky, 2006, Yang & Su, 2007). In the present study, we investigate possibilities of integrating information extracted from Wikipedia and Yago into a coreference resolution system. It has been shown that, even though at earlier stages web knowledge bases might be a source of valuable information (Ponzetto & Strube, 2006), the expansion of such resources inevitably leads to an increase in the amount of noise, making them hardly usable for our application (Ponzetto & Poesio, 2009).

The research line on Wikipedia related to our work is the automatic annotation of terms in a plain text with links to Wikipedia pages. In fact, it is a WSD task because its goal is to link a term in a sentence to the Wikipedia concept that best expresses its sense. Some well-known approaches to this task include (Csomai & Mihalcea, 2008) and (Milne & Witten, 2008). They perform the