Linking Documents to Encyclopedic Knowledge: Using Wikipedia as a Source of Linguistic Evidence

Abstract:

Wikipedia is an online encyclopedia that has grown to become one of the largest online repositories of encyclopedic knowledge, with millions of articles available for a large number of languages. In fact, Wikipedia editions are available for more than 200 languages, with a number of entries varying from a few pages to more than one million articles per language.

In this talk, I will describe the use of Wikipedia as a source of linguistic evidence for natural language processing tasks. In particular, I will show how this online encyclopedia can be used to achieve state-of-the-art results on two text processing tasks: automatic keyword extraction and word sense disambiguation. I will also show how the two methods can be combined into a system able to automatically enrich a text with links to encyclopedic knowledge. Given an input document, the system identifies the important concepts in the text and automatically links these concepts to the corresponding Wikipedia pages. Evaluations of the system showed that the automatic annotations are reliable and hardly distinguishable from manual annotations. Additionally, an evaluation of the system in an educational environment showed that the availability of encyclopedic knowledge within easy reach of a learner can improve both the quality of the knowledge acquired and the time needed to obtain such knowledge.

This is joint work with Andras Csomai.