Babelplagiarism: What can BabelNet do for Cross-language Plagiarism Detection?

Roberto Navigli

Univeristy La Sapienza, Rome, Italy
navigli@di.uniroma1.it

In the first part of the talk, I will present BabelNet, a very large, wide-coverage multilingual semantic network. The resource is automatically constructed by means of a methodology that integrates lexicographic and encyclopedic knowledge from WordNet and Wikipedia. In addition Machine Translation is also applied to enrich the knowledge resource with lexical information for all languages. We present experiments on new and existing gold-standard datasets to show the high quality and coverage of the resource. In a second set of experiments, we show that, when provided with a vast amount of high-quality semantic relations, knowledge-rich word sense disambiguation algorithms compete with state-of-the-art supervised WSD systems in a coarse-grained all-words setting and outperform them on gold-standard domain-specific datasets.

The second part of the talk is devoted to analyzing cases in which BabelNet can be of help in cross-language plagiarism detection. Can a large multilingual semantic network provide hints for detecting plagiarized text? We will see examples of how and when multilingual concepts and disambiguated text can support this task.