Concept Disambiguation in Wikification using Multiple Overlapping Contexts

Mozhgan Saeidi

Abstract

Wikification is a method to automatically enrich a text with links to the encyclopedic knowledge base of Wikipedia. Given an input document, a Wikifier identifies the important entities in the document and links them to the most relevant corresponding Wikipedia pages. Also, it can be considered as automatic tagging of the text. This is beneficial for various applications, like document classification and semantic search. A major challenge is to perform Wikification accurately but also fast enough to process large text documents. An existing approach to speed up coherence-based disambiguation is to divide a large document into chunks, identify a "key entity", a word that carries a strong signal identifying the text's topic, in each chunk, and for each entity, pick the most similar sense to the chosen meaning of the key entity. While this ensures that the cost of Wikification grows linearly with the input size, the partition of the input into disjoint chunks means that the most appropriate key entity to disambiguate a given mention may be in an adjacent chunk. This negatively affects the accuracy of this method.

In this study, our focus is on improving the accuracy of the disambiguation component of the Wikifier, building our system on top of an existing entity recognizer. We demonstrate that using overlapping windows instead of disjoint chunks increases the accuracy of the Wikifier while increasing its computational cost only slightly. In our experimental evaluation using different data sets, we observed an up to 5% higher accuracy using our method than using previous methods. A careful inspection of the entity senses chosen by our method and by the baseline method based on disjoint chunks revealed that our method corrects most of the disambiguation errors made by the baseline method and that these errors arise indeed due to the partition of the input into disjoint chunks in the baseline method. The baseline method never disambiguated an entity correctly that was incorrectly disambiguated by our method.