Abstract:

Wikipedia is a goldmine of information; not just for its many readers, but also for the growing community of researchers who recognize it as a resource of exceptional scale and utility. It represents a vast investment of manual effort and judgment: a huge, constantly evolving tapestry of concepts and relations that is being applied to a host of tasks.

This talk focuses on the process of "wikification"; that is, automatically and judiciously augmenting a plain-text document with pertinent hyperlinks to Wikipedia articles—as though the document were itself a Wikipedia article. I first describe how Wikipedia can be used to determine semantic relatedness between concepts. Then I explain how to wikify documents by exploiting Wikipedia's internal hyperlinks for relational information and their anchor texts as lexical information.

Data mining techniques are used throughout to optimize the models involved. I will discuss applications to knowledge-based information retrieval, topic indexing, document tagging, and document clustering. Some of these perform at human levels. For example, on CiteULike data, automatically extracted tags are competitive with tag sets assigned by the best human taggers, according to a measure of consistency with other human taggers. All this work uses English, but involves no syntactic parsing, so the techniques are language independent.