Project Proposal

Jake Brawer

<2016-04-20 Wed>

1 Proposal

My two great passions are Cognitive Science and Wikipedia. For this project I am endeavoring to merge these two loves. The great thing about Wikipedia is that: 1) it has an intuitive and easily parsed markup language, 2) Articles generally conform to particular style guidelines and therefore are predictable. One incredibly useful feature is that named entities (NEs) in Wikipedia articles are generally hyperlinks to other Wikipedia articles, and therefore are in essence pretagged. However, the NE type is not overtly specified in the hyperlink itself, but implicitly in the content of the associated article. I would therefore like to write a named entity recognition (NER) system that can type extracted NEs by analyzing their own Wikipedia pages. The value of such an algorithm is that it autonomously generate a gigantic corpus of NE, which could be useful for training other NER systems.

Wikipedia is no stranger to information extraction (I.E.). Wu and Weld (2010), for example, sought to extend or create infoboxes in Wikipedia articles, i.e. the relational summary of a given article that are generally present at the beginning of an article. In order to do this, they employed a system called Kylin. Kylin works by identifying Wikipedia articles that share identically structured info boxes (as they usually differ from article to article). Given the attributes in the infoboxes, Kylin tries to identify sentences that contain the corresponding attribute values, and uses these sentences as positive training examples. Kylin trains itself on these sentences in order to identify sentences that contain attributes in test articles.

Yan et al., (2009) designed an unsupervised relation extraction method for Wikipedia articles. The model, given a set of input Wikipedia articles, outputs a list of concept pairs for each article and an associated relation label.

There is a python module aptly called wikipedia which makes extracting information, like hyperlinks, from a given article very simple. I intend to

use that to interface with Wikipedia for the most part. Additionally, If I wish to intelligently extract tags from Wikipedia articles, I'll need some tool for extracting semantics from the text. I think using the Wordnet functions provided by nltk will be good for that.

Given that I've been working on my thesis, I have not gotten to do too much wok on the project as of yet. However, I have played around with the wikipedia module and have successfully tokenized the text of a few articles for practice.

My Three milestones are: 1) By May 3 I would like to have some idea of how to figure out a given article's type. 2) The minimum outcome of the final submission will hopefully be able to type all the the NE in a given article. The ideal outcome would be that my system could be used to extract NEs from a non wikipedia document.

2 References:

- Wu, F., & Weld, D. S. (2010). Open information extraction using Wikipedia. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 118-127). Association for Computational Linguistics.

- Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., & Ishizuka, M. (2009). Unsupervised relation extraction by mining Wikipedia texts using information from the web. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 (pp. 1021-1029). Association for Computational Linguistics.

Emacs 24.5.1 (Org mode 8.3.4)