NOODLE is better than Google*

Bing Bai, Jason Weston and Ronan Collobert
NEC Laboratories America, Princeton, NJ
bbai@nec-labs.com and jasewestonegmail.com and ronan@collobert.com

The use of sophisticated Natural Language Processing (NLP) tools in the field of information retrieval (IR) promises much, but has limited impact in applications to date. Although NLP supported question answering (QA) systems [3] have been successful in domain specific areas [2] and in small document collections such as TREC [3], large scale open-domain QA is still very difficult because the current NLP techniques are still too slow. To the best of our knowledge, no major IR system uses the core NLP tools of syntactic parsing and semantic extraction, e.g. semantic role-labeling, word-sense disambiguation, and anaphora resolution. The best search engines use keyword search to obtain matching documents, and ignore possible syntactic or semantic NLP analysis on either the documents or the query. Without an effective and efficient algorithm, it is impossible to handle a massive database such as the internet.

In this work, we propose NOODLE: an architecture for semantic search. It has the ability to efficiently index and retrieve sentences based on semantic role matching which leverages a neural network architecture similar to SENNA [1] at its core. First, semantic roles are computed *offline* on the large database and this metadata is stored along with the word information necessary for indexing. At query time, semantic roles are predicted for the query *online* and a matching score against each document in the database is computed that incorporates this semantic information. Both *offline* processing and *online* labeling require fast prediction methods, which may explain why this architecture has not been developed before.

In the end, given the indices computed offline, the retrieval time of our system is not much longer than other simple IR models such as the vector space model while the indexing itself is affordable for daily crawling given some reasonable engineering work. More importantly, the neural network framework we employ is general enough to be adapted to other NLP features as well, such as named entity recognition and word-sense disambiguation. We present a preliminary study on Wikipedia with Propbank-based NLP tags where we achieve encouraging results.

References

[1] R. Collobert and J. Weston. Fast semantic extraction using a novel neural network architecture. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 2007.

[2] B.F. Green, A.K. Wolf, C. Chomsky, and K. Laughery. Baseball, an automatic question answerer. In *Proceedings of the Western Joint Computer Conference 19*, pages 219–224, 1961.

[3] E.M. Vorhees. Overview of the trec 2004 question answering track. In *Proceedings of Text REtrieval Conference*, 2004.

Category: web search, neural network
Preference: oral

*Sometimes