

2007 Student Research Symposium

Language Technologies Institute



Language-Independent Set Expansion of Named Entities Using the Web

Richard C. Wang

Abstract:

The work that I would like to present is part of the RADAR project. More specifically, it is part of RADAR's NLP component. The overall goal of this project is to build an intelligent personal desktop agent to assist human on scheduling, processing e-mails, maintaining web pages, and some other daily tasks. The NLP component attempts to understand and process natural text (mostly information extraction from e-mails).

We are trying to improve the NLP component by automatically providing more features for the machine-learned extractors. More specifically, the system we built automatically expands the semantic set of any input named entities in any human language by utilizing the readily available web data. For example, if the inputs are Ford, Nissan, and Toyota, then the outputs would be Buick, Dodge, Mercedes, etc. Alternatively, it can also assist human on expanding knowledge base ontology.

The approach we took uses a wrapper induction technique that 1) automatically builds wrappers based on user inputs and web data, 2) extracts candidate named entities by applying the wrappers on web pages, then 3) builds a graph and performs random graph walk for ranking those extracted entities based on how relevant they are to the inputs.

We constructed a total of 36 evaluation datasets. Each evaluation dataset consists of a publicly well-known and well-defined set of items (i.e. NBA team names, US state names, etc.). The datasets consist of equal number of sets over three languages: English, Chinese, and Japanese. The evaluation procedure is that for each dataset, three items were randomly chosen for expansion, and the resulting ranked list of expanded entities was evaluated using average precision. This process is repeated five times for each dataset, and the mean average precision was computed and reported. The MAP scores were observed to be substantially higher than that of Google Sets (<http://labs.google.com/sets>) on the English evaluation dataset (Google Sets only works for English inputs).

Richard C. Wang and William W. Cohen (2007): Language-Independent Set Expansion of Named Entities using the Web in ICDM-2007.

The above paper is available at <http://rcwang.com/pub/SetExpander.pdf>

A demo of the system described can be found here at <http://rcwang.com/seal>

Here's the abstract of the paper:

Set expansion refers to expanding a given partial set of objects into a more complete set. A well-known example system that does set expansion using the web is Google Sets. In this paper, we propose a novel method for expanding sets of named entities. The approach can be applied to semi-structured documents written in any markup language and in any human language. We present experimental results on 36 benchmark sets in three languages, showing that our system is superior to Google Sets in terms of mean average precision.