INFO256 Project Proposal

Implementation of the XTract Tool in Wordseer

Mosharaf Chowdhury
(21020039)
mosharaf@cs.berkeley.edu
November 5, 2013

Project Description

Natural languages are full of word collocations that frequently co-occur and correspond to arbitrary word usages. They are present in both technical and non-technical textual corpora, and they often have specific significance in individual contexts. Xtract is a statistical tool (i.e., a collection of algorithms) developed for identifying such phrases and statistically justifying their significance. The high-level objective of this project is to implement, in part or whole, the Xtract toolset within Wordseer (a text-analytics platform from UC Berkeley) to replace its default *most-frequent-bigrams-based* phrase extraction mechanism.

Goals

Our detailed goals for this project (and corresponding evaluation strategy) are the following.

1. Implementing the Xtract toolset in a way that can be used with the Wordseer backend. This should be drop-in replacement for the default phrase extraction mechanism in Wordseer.
2. Evaluating the performance of our implementation by comparing against the default Wordseer phrase extraction mechanism. To do this, we will run both implementations through a dataset (which we will identify as part of the project) and compare their precisions head-to-head.
3. Integrating the implementation with the Wordseer frontend (time permitting). This will allow us to see the benefits directly from the frontend, without running anything from command-line.

Members

This is a one-person project. However, the work will be done in close collaboration with and guidance from Aditi Muralidharan, the primary developer of Wordseer.

Resources

1. Wordseer
2. Corpora from NLTK (depending on which domain we want to work on)

Milestones

1. Preparation
    a. Reading the paper on XTract by Smadja. [**November 13**]
2. Implementation
    a. Understanding the Wordseer codebase (backend and frontend). [**November 20**]
    b. Implementing different pieces of the algorithm in the backend. [**November 27**]
    c. Integrating with Wordseer. [**December 11**]
3. Evaluation
    a. Identifying/determining an appropriate domain/dataset; phrases are often domain dependent, so picking a dataset is non-trivial. [**November 27**]
    b. Comparing Xtract performance with that of the default phrase extraction (bigram extraction) mechanism in Wordseer for the chosen dataset/domain. [**December 11**]