The Chinese University of Hong Kong

Department of Computer Science and Engineering

Final Year Project 2007

Mid Term Report

DISCUSSION

The DIStributed Chinese University Sentence Similarity Identification Open Network

Group IK0701

Supervisor: Prof. KING Kuo Chin, Irwin

Members: Sze-To Hei Man, Lire Shelly

Abstract

DISCUSSION is a web application which aims to provide text analysis service to users. By generating comprehensive reports on analyzed text, enable both language learners and native writers to gain deeper understanding on their own writing style. This is a JAVA architecture employing both statistical and dictionary based natural language processing principles. In the current version, DISCUSSION generate report in lexical statistics, part-of-speech tagging and keywords extractions in English. In the future stage, it shall support Chinese analysis also.

Table of Content

Background

In evaluating a piece of writing, there are several aspects to consider. Richness in informations, innovating idea illustrated with concrete arguments, these are the basic qualities a good article of any kind should possess. However, if the writing is written poorly, in terms of both structure and language skills, ideas could be presented in a mess. No one would want to read with difficulties, even if it is good stuff. Proper writing style is essential to be a good writer and a good presenter of one's own opinion.

This logic extends to thesis writing. When most sophisticated ideas are presented by academic writing, possession of excellent language skill is even more crucial for the academics.

The only way leading to improvement is the same old word "practice". But it can be more effective, if one could identify where his weakest is located. In the past, the job of pointing out mistakes and commenting on style are done by language teachers. In the modern age, we try to let the automated computers handle it by AI technology.

We hope to contribute to the further enhancement of our university's integrity in the academy sphere. Combined with CUPIDE, to uphold academic integrity

Objective

In this project, we are going to launch an web application providing integrated text analysis service. Major utilities include the followings:

1. Bilingual system - Chinese and English
2. Lexical Analysis
3. Syntactic Statistics
4. Sementic Analysis
5. Spell Check
6. Grammer Check

This is a JAVA architecture employing both statistical and dictionary based natural language processing principles. In the current version, DISCUSSION generate report in lexical statistics, part-of-speech tagging and keywords extractions in English. In the future stage, it shall support Chinese analysis also.

Development Process

There are two stages in the development schedule.

In the first stage, which we were working on in the current semester, we are to focus on implementation of the backbone functions. We shall develop text analysis service for English in the first semester. Web interface will be in form of a simple query tool (like Google Language Tool). 3 means are supported for user queries.

1. Direct text input via textbox.
2. Feeding web addresses
3. File upload from client's machine

The second stage is basically two major tasks: Implement text analysis for Chinese language and scale up user interface. The simple web interface is to be expanded to a platform with customized user preferences.

Eventually, we hope to merge our product with the CUPIDE system. The ultimate goal is to provide a composite multifunction homework submission platform for both teaching staffs and students.

Previous Work

To address the issue of upholding academic integrity, FYP groups in the previous years have implemented the Chinese University Plagiarism IDentification Engine CUPIDE.

Currently the engine has the following features:

- Support English, Traditional and Simplified Chinese
- Class management web interface
- Electronic homework submissions analysis
- Maintain homework database for institutions
- Support various common document formats
- Generate detailed originality reports
- Highlight suspected plagiarized content

CUPIDE has developed into a rather mature state and is starting to promote its applications to secondary and primary schools and other organizations. There are several editions custom made to suite the need of different user categories.

i. Public Edition

ii. CUHK Edition

    1. for teaching staffs

    2. for students

Study Fields

Text Mining

Text Mining rerefers generally to the process of deriving high quality information from text. High quality information is typically derived through the dividing of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).[1]

Natural language processing (NLP)

Natural language processing (NLP) is a subfield of artificial intelligence and computational linguistics. It studies the problems of automated generation and understanding of natural human languages. Natural language generation systems convert information from computer databases into normal-sounding human language, and natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate.[2]

---

[1] http://en.wikipedia.org/wiki/Text_mining

[2] http://en.wikipedia.org/wiki/Natural_language_processing

System Overview

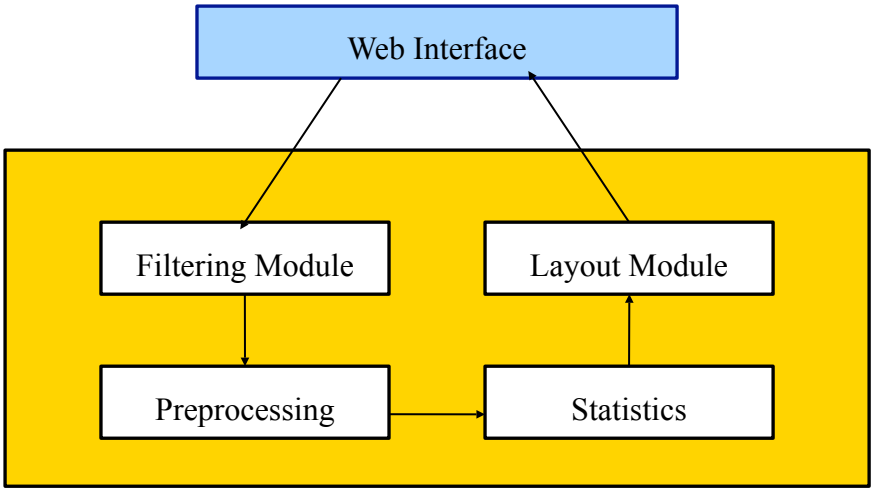

Fig. 1 Structure of the DISCUSSION framework

Web Interface

DISCUSSION is completely web-based. End users could query anywhere and anytime on the engine's website. The interface is written in jsp, enabling the java architecture behind respond to client's request efficiently.

Filtering Module

Besides entering string to the textbox directly, users could submit queries by uploading text files. Most common text files formats are supported. They involves files with extension .html, .pdf, .txt, , The filtering module is responsible for handling I/O of files with different formats.

Preprocessing Module

Perform optional text preprocessing as specified by users.

- Ignore Number

Numbers are ignored if checked. Ignoring numbers would make the analysis more focus on the text-only part of the input by losing its originality. Some important number data could be missed.

- Case Sensitivity

Regard capital letters as small letters or not. Non-case sensitivity analysis is simpler but some special terms may lose their specialness e.g. location name, mathematic terms, date, etc.

- Stopwords Removal

Frequently used words with no semantic significance are called stopwords. For instance, auxiliary verbs, like variations of be. When called, stopwords are truncated according to system's wordlist.

Statistics Module

This is the core module of the engine. Most functions are implemented in this module. Functions are divided into departments by the level of analysis unit. The library OpenNLP is used in the implementation. OpenNLP is an organizational center for open source projects related to natural language processing. Its primary role is to encourage and facilitate the collaboration of researchers and developers on such projects. It hosts a variety of java-based NLP tools which perform sentence detection, tokenization, pos-tagging,

chunking and parsing and named-entity detection.[3] By using these tools, some linguistic works are done and we could focus more in the implementation and layout.

Layout Module

This module is responsible for data visualization. Graphs, charts and sortable table are used extensively to present data in a neat and clear format to feed into the web interface. It is divided into different pages in order to present the data in a more organized and clearer way.

[3] http://www.opennlp.org/

Web Interface layout

fig.2 DISCUSSION front page. User input textbox.

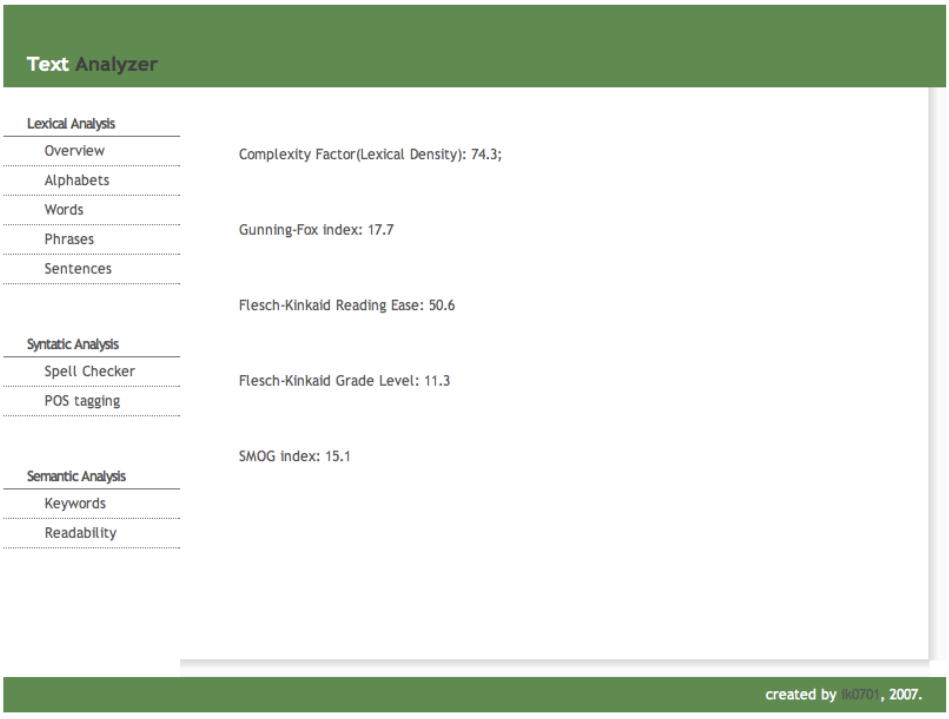

fig.3 Readability Report

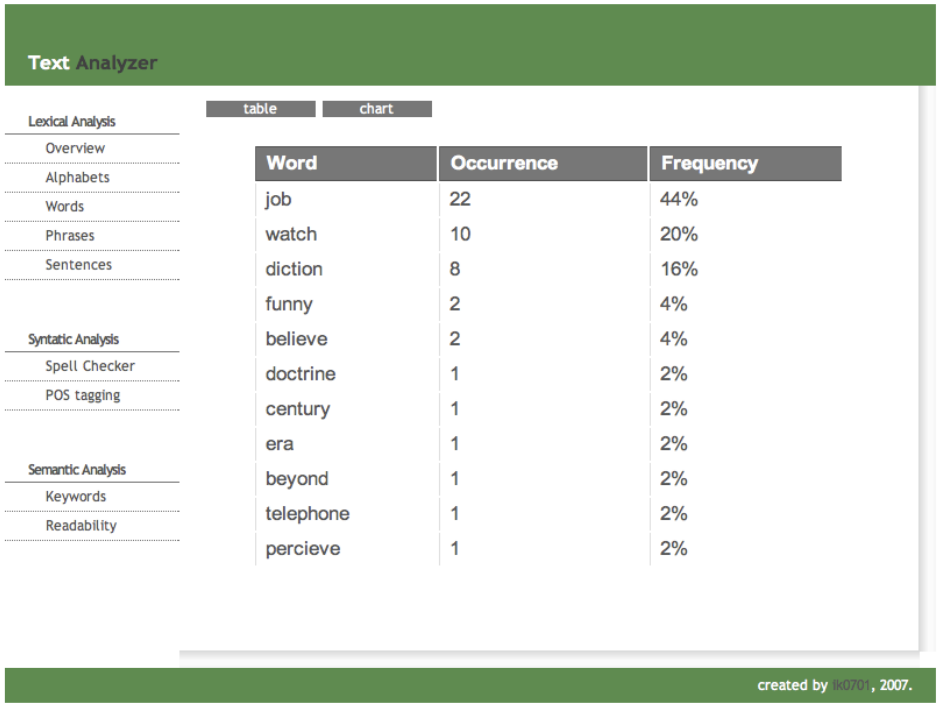

fig.4 Top words analysis - table format

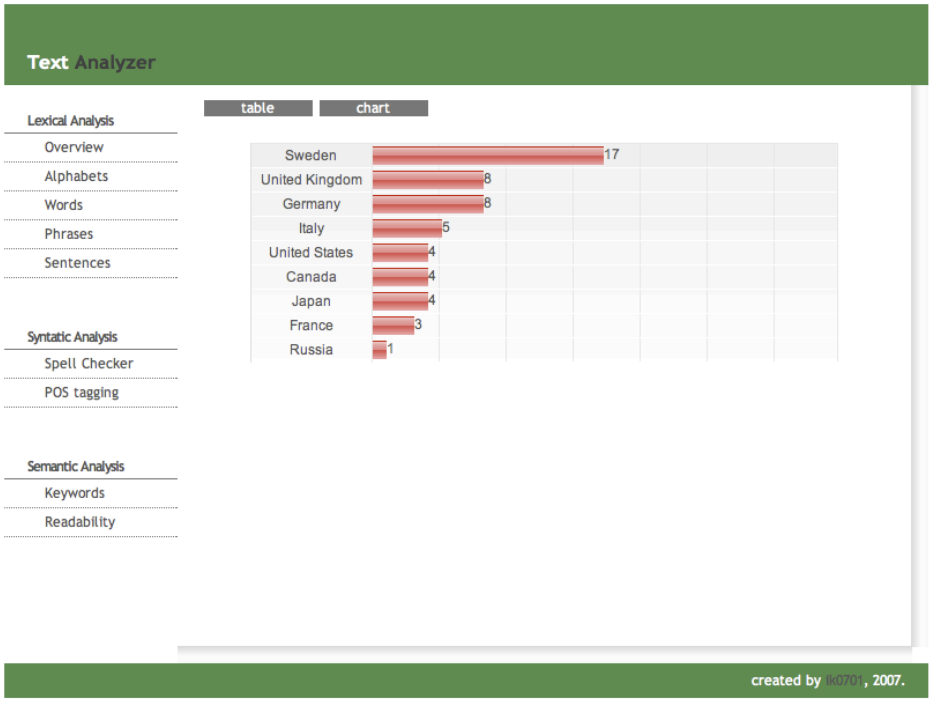

fig.5 Finding country names in DISCUSSION - bar chart format

Functionalities

Much information are extracted from interpreted text. According to the analysis aspects, functions are grouped into 3 categories, namely, lexical, syntactic and sementic analysis.

I. ***Lexical Analysis***

In linguistics, the lexicon of a language is its vocabulary, including its words and expressions. More formally, it is a language's inventory of lexemes.[4]

General statistics

Essential statistics extract. Provide an overview of basics characteristics of the queried text.

Alphabets

Count the number of alphabets, with space and without space. A list of the number of alphabets would be given out.

Words

Basic statistics: word count, unique word count etc. Unique word count would show the number of distinct words in the input.

Top Words

A list of words ranking with frequency. It can be sorted by alphabetical order, frequency and number of occurence. Tops words list can help in observing the most important words in the text and its main idea.

[4] http://en.wikipedia.org/wiki/Lexicon

Phrases

Count occurrence of phrases with difference lengths, typically 2 to 5. It has similar function as top words list but in units of phrases.

Sentences

Basic sentence counting, sentence length counting etc. Individual sentences are analyzed to get a more detail report of the input text.

II. *Syntactic Analysis*

Syntactics refer to the relations between signs in formal structure. It has similar meaning to grammatical structure in natural language.

A. Spell Checker

Indicates misspelled words and suggest corrections. Statistics of misspell frequency. A toy spelling corrector implemented by Peter Norvig is used.[5] It is trained by a text file and achieves 80% or 90% accuracy at a processing speed of at least 10 words per second.

B. Part-of-Speech Tagging

Tag every word with its part of speech. Statistics of part-of-speech occurrence. It is done by using the class PosTagger in the OpenNLP package. A dictionary and model file provided by OpenNLP are used for training. Although it is not very fast but the accuracy is quite high.

[5] http://www.norvig.com/spell-correct.html/

III. ***Semantic Analysis***

Semantics refers to the relation between signs and the things they refer to, that is their denotation.

A. Keywords

Extract important informations from text. Potential keyword are numbers, dates and times, names of people, organization and places etc. The class NameFinder in OpenNLP is used. It provides name tags for a sequence of tokens by using training models of different kinds of names.

B. Readability

Give various readability indexes. i.e.

- Gunning Fog Index
  - It is a test designed to measure the readability of a sample of English writing. The resulting number is an indication of the number of years of formal education that a person requires in order to easily understand the text on the first reading.[6]
- Automated Readability Index
  - It is a readability test designed to gauge the understandability of a text. Like the Flesch-Kincaid Grade Level, Gunning-Fog Index, SMOG Index and Coleman-Liau Index, its output is an approximate

[6] http://en.wikipedia.org/wiki/Gunning_fog_index/

representation of the U.S. grade level needed to comprehend the text.[7]

- SMOG Simple Measure of Gobbledygook
  - It is a readability formula that estimates the years of education needed to understand a piece of writing. It is widely used, particularly for checking health messages. It can substitute the Gunning-Fog Index.[8]
- Flesch-Kincaid Readability Tests
  - They are readability tests designed to indicate how difficult a reading passage is to understand. There are two tests, the Flesch Reading Ease, and the Flesch-Kincaid Grade Level. In the Flesch Reading Ease test, higher scores indicate material that is easier to read. The Flesch-Kincaid Grade Level translates the 0-100 score to a U.S. grade level.[9]
- Coleman-Liau Index
  - It works like the Flesch-Kincaid Grade Level, Gunning-Fog Index, SMOG Index, and Automated Readability Index which tells the U.S. grade level thought necessary to comprehend the text.[10]

[7] http://en.wikipedia.org/wiki/Automated_Readability_Index/

[8] http://en.wikipedia.org/wiki/SMOG_%28Simple_Measure_Of_Gobbledygook%29/

[9] http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test/

[10] http://en.wikipedia.org/wiki/Coleman-Liau_Index/

Conclusion

Natural language processing, being one of the toughest problem in computer science, is a challenging task for our group. In the developing process, we have came across many obstacles, which constant reference reviews and careful deliberations were needed. These consumption of time is way far more than expected. In the end, we are capable of implementing some basic functionalities of the DISCUSSION system, though, at the same time we found ourselves lagging behind the planned development schedule. For instance, intended.We hope to catch up in the next semester. After exploring large amount of related works in this field and gained understandings, we believe that DISCUSSION will become more copious.

Reference

OpenNLP http://opennlp.sourceforge.net/

MBT: Memory Based Tagging demo http://ilk.uvt.nl/~zavrel/tagtest.html/

Topicalizer http://www.topicalizer.com/

Textalyzer http://textalyser.net/

Natural Language Toolkit NLTK http://nltk.sourceforge.net/index.php/Main_Page

CMU Text Learning Group http://www.cs.cmu.edu/~TextLearning/

WordNet http://wordnet.princeton.edu/

TextArc http://www.textarc.org/

Lemur Toolkit http://www.lemurproject.org/

White Smoke - Grammer Checker http://www.whitesmoke.com/

Grammar Expert Plus http://www.wintertree-software.com/app/gramxp/

Bakeoff http://www.china-language.gov.cn/bakeoff08/

Jython http://www.jython.org/Project/index.html/