ISA 562100 NLP Lab -- Project Description

Jason S. Chang

Class Project

The purpose of the research project is for the students to learn how to formulate a simple natural language problem/task/application and to experience how to solve it using methods, algorithms and techniques taught in class. The students will conduct experimental evaluation on an interesting dataset and will analyze the obtained results. Students are encouraged to identify new problems/tasks/applications, however we will also provide them with a sample of topics. Building a demo is a must. We strongly encourage the students to work in groups of 2 people.

Project Timeline

2014/1125: Term Project Announcement, EM
2014/1202: a-Linggle: Linguistic Search Engine for Academic Writing
(discussion about how to design map/reduce)
2014/1209: Writing proposals
2014/1216: a-Linggle (2) / Proposal due / Introduction to Flask
https://realpython.com/blog/python/flask-by-example-part-1-project-setup/
2014/1223: Interim oral report
2014/1230: Interim oral report
2015/0106: Final report

Sample Project

- a-MOVER (Sentence classifier)
  - See the following paper:
    https://www.researchgate.net/publication/3230281_Mover_a_machine_learning_tool_to_assist_in_the_reading_and_writing_of_technical_papers
  - Use the data available in
    http://archive.ics.uci.edu/ml/machine-learning-databases/00311/
- a-Conc (Concordancer for academic writing)
  - Distant Supervision and Word Alignment
  - Query expansion using translation based paraphrasing
- a-Checker (Concordancer for academic writing)
  - E.g., * … propose a new to find … -> … propose a new for finding …)
  - May focusing on verb selection errors (with prepositions, article)
- a-Linggle (linguistic search engine for academic writing)
  - May focusing on verb selection errors (with prepositions, article)
- a-GDEX (linguistic search engine for academic writing)
  - May focusing on verb selection errors (with prepositions, article)

Natural Language Processing Lab
Term Project Proposal

Jim Chang and Joseph Yen

Grammatical Error Correction: Wrong Lexical Choice of a Verb

Abstract

(350-500 words)

The notion of collocation has been widely discussed in the field of language teaching for decades. It has been shown that collocation is important in helping language learners achieve native-like fluency. In the field of English for Academic Purpose, more and more researchers are also recognizing this important feature in academic writing. It is often argued that collocation can influence the effectiveness of a piece of writing and the lack of such knowledge might cause cumulative loss of precision.

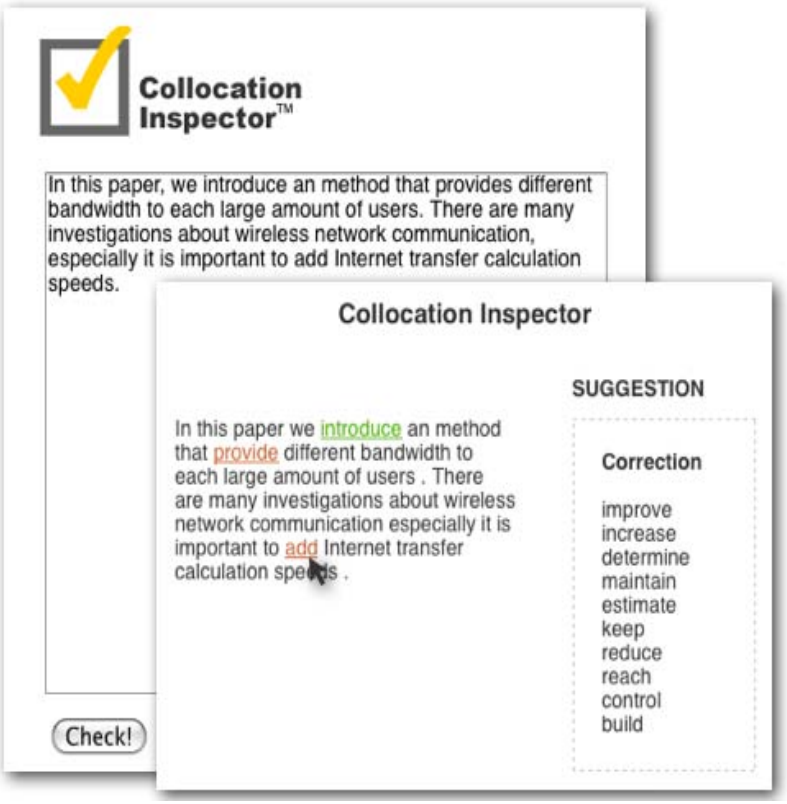
To tackle such word usage problems, traditional language technology often employs a database of the learners' common errors that are manually tagged by teachers or specialists (e.g. Shei and Pain, 2000; Liu, 2002). Such system then identifies errors via string or pattern matching and offer only pre-stored suggestions. Compiling the database is time-consuming and not easily maintainable, and the usefulness is limited by the manual collection of pre-stored suggestions. Therefore, it is beneficial if a system can mainly use untagged data from a corpus containing correct language usages rather than the error-tagged data from a learner corpus. A large corpus of correct language usages is more readily available and useful than a small, labeled corpus of incorrect language usages.

For this suggestion task, the large corpus not only provides us with a rich set of common col- locations but also provides the context within which these collocations appear. Intuitively, we can take account of such context of collocation to generate more suitable suggestions. Contextual information in this sense often entails more linguistic clues to provide suggestions within sentences or paragraph. However, the contextual information is messy and complex and thus has long been overlooked or ignored. To date, most fashionable suggestion methods still rely upon the linguistic components within collocations as well as the linguistic relationship between misused words and their correct counterparts (Chang et al., 2008; Liu, 2009).

Goals of the project

We propose a machine learning approach to implementing an online collocation writing-assistant. We use a data-driven classifier to provide collocation suggestions to improve word choices, based on the result of classification. The system generates and ranks suggestions to assist learners' collocation usages in their academic writing with satisfactory results.

Problem Statement: Given a sentence *S* written by a learner and a reference corpus *RC*, our goal is to output a set of most probable suggestion candidates $c_1, c_2, \ldots, c_m$. For this, we train a classifier *MC* to map the context (represented as feature set $f_1, f_2, \ldots, f_n$) of each sentence in *RC* to the collocations. At run-time, we predict these collocations for *S* as suggestions.



Detailed plan of activities *(Include a timeline and assignments of responsibilities.)*

(1) Academic Collocation Checker Training Procedures

1.1 Parse sentences and extract collocation as training data

1.2 Select features for machine learning

1.3 Train a classifier

(2) Automatic Collocation Suggestion at Run-time

2.1 Parse the input sentence and extract VO relations

2.2 Select features for classification

2.3 Run the classifier to suggest alternative verbs for the verb in each VO

2.4 Obtain (verb, probability) pairs from the classifier

2.5 Display verbs in descending order of probability

Data and resources

1. British National Corpus
2. Stanford Parser (http://nlp.stanford.edu/software/lex-parser.shtml)
    - http://nlp.stanford.edu:8080/parser/

References

Liu, Anne Li-E., David Wible, and Nai-Lung Tsao. "Automated suggestions for miscollocations." *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2009.

Chang, Yu-Chia, et al. "An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology." *Computer Assisted Language Learning* 21.3 (2008): 283-299.

Wible, David, et al. "A Web-based EFL writing environment: integrating information for learners, teachers, and researchers." *Computers & Education* 37.3 (2001): 297-315.

MH Chen, Factors and Analysis of Common Miscollocations of College Students in Taiwan. 2011.

Wu, Jian-Cheng, et al. "Automatic collocation suggestion in academic writing." *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, 2010.