Project Milestone 1

Xiaojie Zhu XXZ180012

Bo Jin JXB180009

Hongzheng Wang HXW180004

Paper 1:

1. **Citation:**

Tsur, Oren & Rappoport, Ari. (2007). Using Classifier Features for Studying the Effect of Native Language on the Choice of Written Second Language Words. 10.3115/1642025.1642027.

2. **Task:**

Tsur and Rappoport (2007) applied machine learning techniques to study language transfer, a major topic in the theory of Second Language Acquisition (SLA). They used an SVM for the problem of native language classification and showed that native language has a strong effect on the word choice of people writing in a second language. This partly explains the differences between the writings by native speakers and foreign language speakers.

3. **Data:**

Tsur and Rappoport used the International Corpus of Learner English (ICLE) as the corpus used in the experiments in this paper. They worked on 5 sub-corpora, each containing 238 randomly selected essays by native speakers of the following languages: Bulgarian, Czech, French, Russian and Spanish, and the essays are of two types: argumentative essays and literature examination papers.

4. **Approach:**

Tsur and Rappoport first implemented a naive baseline classifier to test the unigram baseline accuracy and compare it with the bigram baseline results. They then chose the 200 most frequent character bi-grams in the corpus, and used a vector of the same dimension. Each vector entry contained the normalized frequency of one of the bi-grams. They then used tri-gram frequencies as features and repeated the same experiment with the top 200 trigrams.

Further, they trimmed the function words and address the function words bias, content bias, and suffix bias to compare the effects of each different features on the choice of words when writing in a second language. Finally, they also ran the experiment on a different corpus replacing the French and the Spanish sub-corpora

by the Dutch and Italian ones, introducing a new Roman language and a new Germanic language to the corpus, to see if the results also fits for other languages.

5. **Evaluation:**

Tsur and Rappoport used the multi-class SVM and obtained 46.78% accuracy on the unigram baseline, which is more than twice the random baseline accuracy, and is in accordance with the bigram results. They then used a multiclass SVM in a 10-fold cross validation manner to examine the accuracy of the bigram model and achieved 65.60% accuracy with standard deviation of 3.99. They then repeated the same experiments with tri-gram frequencies as features, and yielded 59.67% of accuracy, which is 40% higher than the expected baseline and 15% higher than the uni-grams baseline.

They further removed all function words from the corpus and ran the experiment once again, which achieved an accuracy of 62.92% in the 10-fold cross validation test. The test on the content bias and suffix bias indicated that content bias and suffix bias existed in the corpus add only a minor effect on the SVM classification. Finally, they also ran the experiment on a different corpus replacing the French and the Spanish sub-corpora by the Dutch and Italian ones and obtained 64.66% accuracy, essentially the same as in the original 5-language setting.

Paper 2:

1. **Citation:**

Michael Gamon, Using Mostly Native Data to Correct Errors in Learners' Writing: A Meta-Classifier Approach. HLT'10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Page 163 - 171

2. **Task:**

Michael Gamon proposed method for error correction for non-native speakers of English. By firstly compare a language model and error specific classifiers (all trained on large English corpora) with respect to their performance in error detection and correction, and then combine the language model and the classifiers in a meta-classification approach by combining evidence from the classifiers and the language model as input features to the metaclassifier, correction can be achieved. This is relatable to the project we propose.

3. **Data:**

Cambridge University Press Learners' Corpus (CLC). The version of CLC that we have licensed currently contains a total of 20 million words from learner English essays written as part of one of Cambridge's English Language Proficiency Tests (ESOL) – at all proficiency levels

4. **Approach:**

First compare a language model and error specific classifiers (all trained on large English corpora) with respect to their performance in error detection and correction, and then combine the language model and the classifiers in a meta-classification approach by combining evidence from the classifiers and the language model as input features to the metaclassifier. The learning is supervised learning. The learning algorithm is maximum entropy classifier (Rathnaparki 1997) for articles and for prepositions. Features include contextual features from a window of six tokens to the right and left, such as lexical features (word), part-of-speech tags, and a handful of "custom features", for example lexical head of governing VP or governed NP.

5. **Evaluation:**

Evaluation involves running the meta-classifier system on the preposition and article test sets described in above and calculate precision and recall. So evaluation metric is simply precision, recall, and f-measures. Intrinsic evaluation of word vectors is the evaluation of a set of word vectors generated by an embedding technique on specific intermediate subtasks. Extrinsic evaluation of word vectors is the evaluation of a set of word vectors generated by an embedding technique on the real task at hand. The baseline to compare against is CLC annotation library.

Paper 3

1. **Citation**

Azab, Mahmoud; Salama, Ahmed; Oflazer, Kemal; Shima, Hideki; Araki, Jun; Mitamura, Teruko (2018): An NLP-based Reading Tool for Aiding Non-native English Readers. figshare. Journal contribution.

2. **Task**

Mahmoud Azab , Ahmed Salama, Kemal Oflazer, Hideki Shima, Jun Araki and Teruko Mitamura builds a text-reading tool, a web-based tool, to help non-native English speakers overcame languages barriers. In this paper, their task is to enable users to interact with the texts information anytime and anywhere.

3. **Dataset**

The dataset is WordNet, which is a broad-coverage machine-readable dictionary of English.

4. Approach

They implemented a web-based browser application, which can interact with the text either clicking a word or select any parts of the text to get the lexical information, syntactic information and so on. The learning is supervised learning. The tool is based on a client-server software architecture, with the UIMA-framework being used for both annotations and querying and it can be integrated into an e-book reader when users reading can use this tool.

5. Evaluation

The evaluation is the selection based on their library and test some short documents for intrinsic evaluation and report their recall and precision. For extrinsic evaluation of this tool, they would have a group of non-native English speakers to use and record the experience and then evaluate it. The baseline they compare against is the user experience and the accuracy of using WordNet.