Word Sense Disambiguation for Vocabulary Learning

Anagha Kulkarni, Michael Heilman, Maxine Eskenazi and Jamie Callan

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave,
Pittsburgh, PA, USA 15213
{anaghak,mheilman,max,callan}@cs.cmu.edu

Abstract. Words with multiple meanings are a phenomenon inherent to any natural language. In this work, we study the effects of such lexical ambiguities on second language vocabulary learning. We demonstrate that machine learning algorithms for word sense disambiguation can induce classifiers that exhibit high accuracy at the task of disambiguating homonyms (words with multiple distinct meanings). Results from a user study that compared two versions of a vocabulary tutoring system, one that applied word sense disambiguation to support learning and another that did not, support rejection of the null hypothesis that learning outcomes with and without word sense disambiguation are equivalent, with a *p*-value of 0.001. To our knowledge this is the first work that investigates the efficacy of word sense disambiguation for facilitating second language vocabulary learning.

Keywords: Vocabulary Learning, Word Sense Disambiguation, English as a Second Language, Computer Assisted Language Learning

1 Introduction

Learning vocabulary is central to learning any language. Learning a new word implies associating the word with the various meanings it can convey. Consequently it is helpful to think of acquiring vocabulary as a task of learning word-meaning pairs, such as, {(word_1, meaning_1), (word_1, meaning_2), (word_2, meaning_1)} rather than a task of learning words {(word_1), (word_2)}. This approach is, of course, more relevant for some words than others. Many words can convey only a single meaning. However for many other words that is not the case and such words are termed as ambiguous words. It is important for an intelligent tutoring system designed to assist English as a Second Language (ESL) students to improve their English vocabulary, to operate at the level of the word-meaning pairs being learned and not just the words being learned, for several reasons. The most important reason is to be able to assess learning of the particular meanings of a word that the student was exposed to. The second reason is to personalize and adapt the tutoring material in order to expose the student to all or a particular set of meanings of a word. These observations motivate the study of word meaning/sense disambiguation (WSD) for supporting vocabulary learning in a tutoring system.

2 Background

For this study, we extended an existing tutoring system for ESL vocabulary, which is described below in Section 2.1. Sections 2.2 and 2.3 provide a background about the phenomenon of word sense ambiguity and its aspects relevant to this study.

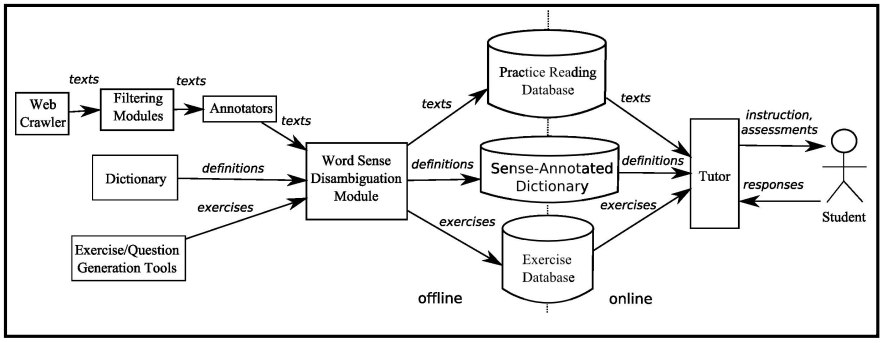

Fig. 1. REAP System Diagram

2.1 REAP Tutoring System

The REAP tutoring system [1] assists ESL students in improving their vocabulary. Its approach is that of context-based teaching, where the word, its meaning and its usage are demonstrated by exposing the student to the word in a natural text (the practice text) instead of in isolation. For each grade level a human teacher prescribes a set of new words (focus words) that are appropriate for that level. The REAP system personalizes instruction for each student by allowing the student to indicate through self-assessments the words that he or she is familiar with, which modifies word priorities. In each instructional cycle, the student chooses from four different practice texts, each containing two or more words from the student's personalized word list. Once the student has finished reading the text, he or she answers practice questions that are based on the words that he or she was exposed to in the text.

The practice texts containing focus words are selected from the World Wide Web (WWW). As such, REAP brings together the two sub-disciplines of language technologies, Information Retrieval (IR) and Computer Assisted Language Learning (CALL). The quality of the practice text is controlled by selecting only those documents from the WWW that pass through various automatic filters implemented in the REAP system, for instance, text-quality filter, reading-level filter and document length filter. REAP's system diagram is shown in Figure 1. To maximize the time on task, documents containing multiple focus-words are preferred. Students can also indicate preferences for 16 general topics including Arts, Science, and Sports, which are taken into consideration by the system while choosing the practice texts for each student. A machine-readable version of an ESL dictionary, the Cambridge Advanced Learners Dictionary (CALD) [3] is integrated into REAP. Students can use CALD while they read a document, to lookup definitions and examples of a word.

2.2 Lexical ambiguity in language

Most languages have words with multiple meanings. Zipf [4] proposed several empirical laws based on the *Principle of Least Effort* to explain some of the phenomena observed in languages. One law proposes that word sense ambiguity arises as a result of two competing forces, both of which try to follow the principle of least effort. On one side, the effort of the speaker of the language will be minimized by using a small vocabulary, that is, by using few words that are capable of conveying many meanings. On the other side, the listener's effort will be minimized by using as many distinct words, in terms of meaning, as possible. Zipf provides empirical evidence for his theory. However, the conclusion most pertinent to this study is that his theory formalizes the common belief that human language vocabularies are a mix of ambiguous and unambiguous words.

2.3 Polysemes and Homonyms

Ambiguous words can be categorized as *polysemes* or *homonyms*. Polysemes are words that can convey multiple related meanings, whereas, homonyms are words that can convey multiple distinct meanings. For example, the word *branch* as a noun has the following definitions listed in CALD that are all related.

1. one of the parts of a tree that grows out from the main trunk and has leaves, flowers or fruit on it
2. a part of something larger
3. one of the offices or groups that form part of a large business organization
4. a part of a river or road that leaves the main part

whereas the following definitions for the word *bark* in the noun form convey clearly distinct meanings.

1. the loud, rough noise that a dog and some other animals make
2. the hard outer covering of a tree

In this study we concentrate on homonyms for two reasons. First, distinguishing between related senses of a word is a highly subjective task. Several studies [5, 6, 7] have shown that the agreement between human annotators is very low on this task. Second, we believe that ESL students can transfer their knowledge about one sense of a word to another related sense of the word without much difficulty, especially in a context-based learning setup. However, we hypothesize that learners are not able to do so for homonyms, and thus assistance should improve learning.

Given this background the two objectives of this work can be stated as:

1. to demonstrate that automatic disambiguation of homonyms that occur in web-pages is possible and,
2. to show that such methods can be applied in a tutoring environment to produce a positive effect on ESL vocabulary learning.

3 Word Sense Disambiguation Methodology

WSD is a well-established research area in the field of natural language processing [8]. Here we concentrate on word sense disambiguation of homonyms that occur in web-pages. For the purposes of this paper we will refer to the task of homonym disambiguation as WSD. This task can be structured as a supervised learning problem where the availability of an annotated training-set is assumed or as an unsupervised learning problem where a training-set is not available. In the supervised learning setting WSD becomes a text classification problem and in the unsupervised setting it is often referred as the word sense discrimination task [15, 16]. The unsupervised framework has the advantage of not requiring a training-set. We experimented with both approaches. However, models learned using supervised methods were consistently more accurate than models learned using unsupervised techniques, thus we focus on the supervised methods in this paper. The decision to use supervised learning techniques was motivated by the need to minimize the potential effects of classification errors on student learning.

The supervised WSD task can be stated formally as follows. The following are given: 1) a homonym h, 2) the set M of distinct meanings that h can convey, and 3) a training-set T that consists of (i, k) pairs where i is a portion of text containing h, and k is the meaning that i conveys. The goal is to learn the best classification model $\hat{C}$ using T for h. A best classification model $\hat{C}$ would be one that generalizes well, that is, it not only performs classification with high accuracy on T but also on the test-set S, which consists of instances of h that were not used for inducing the model. Note that the phrase 'the portion of text' used above is a generic phrase that can refer to a phrase, a sentence, a paragraph, or an entire document containing h. The task of learning a text classification model is typically divided into two broad phases – feature extraction and classification algorithm. Section 3.1 describes the features that were used in this study. We used the Weka [10] implementation of two standard machine learning algorithms, namely, Multinomial Naïve Bayes (MNB) [11] and Support Vector Machines (SVM) [12]. Section 3.1 briefly describes these algorithms.

3.1 Features types and Classification algorithms

Loosely speaking, classification models generalize by learning patterns of co-occurrences from the training data. For instance, a classification model for the homonym *bark* might learn from the training examples that if the word *dog* occurs in the vicinity of the word *bark* then it is most likely that the first meaning of *bark* related to dogs, is at work in this instance. In general, identifying multiple such indicators or features from the training data can lead to learning a robust classification model, assuming that the task under consideration lends itself to such abstractions.

We use two types of features, namely, unigrams (UNI) which are lexical features, and part-of-speech-bigrams (POS-BI) which are lexico-syntactic features. [13] shows that using such a combination of features is effective for supervised WSD. Unigrams are the unique words that occur in the training-set instances. However, only those unigrams that occurred within some window $(-w, +w)$ of the homonym are considered. The intuition behind this approach is that a word's local context is more

likely to encode information about the word and its sense than distant context. The window size was one of the parameters that we varied in our experiments (Section 4). Closed-class words such as articles, prepositions and numerals were not included because they rarely help distinguish between the different meanings of a homonym Unigrams that occur only once were also discarded. Generating the part-of-speech-bigrams was a two step process. In the first step, each word which was within a window of five words on the left and right of the homonym in each training-set instance was annotated with its part-of-speech tag, such as, noun, verb, adjective, using the Stanford Part-Of-Speech tagger[1]. Given this sequence of part-of-speech tags, unique pairs of consecutive part-of-speech tags are extracted in the second step. POS-BIs capture the syntactic information about the neighborhood of the homonym, and provide another level of abstraction. Generating other features such as lexical bigrams, trigrams or pos-trigrams, is possible. However, the available evidence for these features becomes very sparse. Intuitively, the number of occurrences in the training-set of the trigram "loud dog bark" would be much less than that of "loud", "dog", and "bark" individually.

The Multinomial Naïve Bayes algorithm [11, 14] is a variation of the classification algorithm based on the Bayes' Rule. MNB makes two assumptions: i) the features are independent of each other given the class (the sense of the word in our case), and ii) the class conditional distribution of the features is a multinomial distribution. Support Vector Machines [12] identify a classification boundary that is maximally separated from all the classes (again, the senses of the homonym, in our case). MNB and SVM are well-known, frequently used, and frequently effective; however, they make very different assumptions about the data. As we will see, the assumptions made by MNB are empirically found to be more appropriate for our data.

4 Experiments with WSD Approaches

We focus on 30 homonyms in this study. The different morphological forms of these homonyms were treated as the same word type. For instance, along with the root form of the homonym *issue*, the derived form *issues,* was also analyzed. The list of 30 words in their root form is given in Table 1. Following is the description of the training-set generation process.

The definitions of a word provided by an ESL dictionary, the Cambridge Advanced Learners Dictionary (CALD), were manually grouped based on the relatedness of the meaning that they convey. An example, for the homonym *issue,* is shown below:

Group 1

1. a subject or problem which people are thinking and talking about

Group 2

2. a set of newspapers or magazines published at the same time or a single copy of a newspaper or magazine

Group 3

3. An issue of shares is when a company gives people the chance to buy part of it or gives extra shares to people who already own some.
4. to produce or provide something official

[1] http://nlp.stanford.edu/software/tagger.shtml

The third column in Table 1 specifies the number of definition groups (senses) for each of the 30 words that were used in this study. These definition groups were used as the sense-inventory for the coding task, which is described next. The training-set was created for each word by annotating occurrences of the word in web-pages, using the word's sense-inventory. This annotation task was performed by an independent group of coders from the Qualitative Data Analysis Program[2] at the University of Pittsburgh. In the initial round, four coders annotated 100 documents. (We use the words 'document' and 'web-page' interchangeably.) The pair of coders with best inter-coder agreement (Cohen's *kappa* = 0.88) was chosen to annotate the remaining documents. To expedite the annotation task, both coders annotated different sets of documents. A "spot-check" process was implemented by periodically providing a subset of the documents to both the coders for annotation. The inter-coder agreement for the entire training-set, in terms of Cohen's *kappa*, based on spot-checks, was 0.92. These high kappa values provide empirical evidence that the task of distinguishing between distinct senses is much less subjective than the task of distinguishing between fine-grained senses, and thus can be automated much more effectively. The annotated dataset thus generated was used in the experiments described below.

For each word, classification models were learned using the two machine learning algorithms described in Section 3.2. Further more, 46 different window sizes (10 through 100, in steps of 5), for the unigram feature extraction task were experimented with. 10-fold cross-validation [9] was used to compare the different classification models learned. The best learning algorithm for each word and the best window size is specified in the second last and the last columns of the Table 1.[3] Multinomial Naïve Bayes algorithm was the best classification algorithm for 22 of the 30 homonyms. The average best window size was 25 (-25, +25). The best accuracy values for each word are compared with the baseline accuracy given in the column 4. The baseline accuracy is computed by assigning labels to any given instance of the word with the *most frequent sense* of the word in the training-set. As the table shows, for some words the baseline accuracy is quite high (e.g., factor, function) indicating that one sense is extremely dominant. This can happen when all or a majority of the instances of the word in the training-set belong to the same topic, such as, science, or arts. Cohen's *kappa*, reported in column 6, indicates the agreement between the gold standard and the predicted senses. Table 1 is sorted on the *kappa* values.

5 User Study and Discussion

A user study was conducted at the English Language Institute (ELI)[4], at the University of Pittsburgh, to test the effects of WSD on ESL vocabulary learning. A total of 56 students from the ELI Reading 4 and Reading 5 classes (respectively upper intermediate and advanced ESL students) participated in the study. The Reading 4

[2] http://www.qdap.pitt.edu/

[3] Note that in this setting it is practical to use different learning algorithms and window sizes for each word, if that yields the best accuracy.

[4] http://www.eli.pitt.edu/

group consisted of 39 students, and the Reading 5 group consisted of 18 students. The pre-test consisted of 30 self-assessment questions similar to the Yes/No test [17], one for each homonym, where each student was asked to indicate if he or she knew the word. The study was conducted for duration of eight consecutive weeks; one session per reading level per week. The pre-test was conducted, during the first session, and lasted approximately 10 minutes. The practice reading activity started during the same session and continued for the following seven consecutive weeks, one session per week. The post-test was conducted during the eighth and final session. It consisted of cloze questions for each of the <word, sense> pairs that the student was exposed to during the practice reading sessions. These cloze questions were manually created by the teachers at the ELI. Out of the 56 students 47 students attended the final session, the post-test. The following analysis is based only on these 47 students. The experimental group that used the WSD-equipped REAP consisted of 24 students and the control group consisted of 23 students. General ESL proficiency was measured by Michigan Test of English Language Proficiency (MTELP) scores.

Table 1. Summary of Supervised classification models for 30 homonyms.

	Homonym	Number of senses	Baseline Accuracy (%)	Best Accuracy (%)	Cohen's Kappa	Classification Algorithm	Window Size (*w*)
1	panel	2	84.92	99.82	0.993	MNB	25
2	transmission	2	51.91	99.15	0.983	MNB	70
3	procedure	2	82.04	99.40	0.980	MNB	85
4	foundation	3	79.17	98.81	0.965	MNB	85
5	principal	2	64.39	98.05	0.957	SVM	40
6	bond	2	81.78	98.67	0.956	SVM	70
7	aid	2	59.76	97.71	0.952	SVM	85
8	tape	2	75.16	98.14	0.951	MNB	40
9	monitor	2	84.36	98.58	0.947	MNB	85
10	code	2	66.18	97.10	0.936	MNB	85
11	volume	3	51.00	96.00	0.934	MNB	85
12	suspend	2	81.48	97.53	0.919	MNB	40
13	contract	3	83.67	97.73	0.919	MNB	40
14	qualify	3	79.81	97.12	0.909	MNB	70
15	major	3	90.24	98.32	0.904	MNB	40
16	conceive	2	80.92	96.95	0.898	SVM	70
17	pose	3	58.26	94.78	0.893	MNB	25
18	trigger	2	59.40	94.33	0.883	SVM	25
19	brief	3	75.81	95.70	0.883	SVM	10
20	parallel	2	53.70	94.14	0.882	MNB	85
21	supplement	2	73.18	95.45	0.882	MNB	70
22	channel	2	53.25	93.49	0.869	MNB	10
23	depress	2	60.66	93.44	0.862	MNB	40
24	manual	2	68.80	93.59	0.850	SVM	10
25	factor	2	91.24	97.72	0.848	MNB	85
26	shift	3	70.71	92.55	0.837	MNB	70
27	function	2	90.84	97.01	0.830	MNB	55
28	issue	3	80.90	92.96	0.767	MNB	85
29	complex	3	58.51	86.86	0.735	MNB	70
30	appreciate	2	68.63	86.27	0.690	SVM	40

The word sense disambiguation models learned for the 30 homonyms, that are described in Section 4, were integrated into the REAP system to support vocabulary learning. This modified the system in two main ways. First, during the reading session whenever a student looked up one of the 30 homonyms in the integrated dictionary, the order of the definitions was adjusted to show the most appropriate definition for that particular document at the top of the list. Prior research such as [4] motivates this by observing that dictionary users often stop reading the entry for a word after reading only the top few definitions, irrespective of whether those provide them the correct definition for the given usage of the word.

The second change improves the quality of the multiple-choice practice questions that follow each reading. Each question requires the student to select one among five alternative definitions of a word that he or she was exposed to in the reading. Generating these five alternative definitions (four distractors and one correct definition) is straightforward for words with single sense. However, for homonyms, without WSD it is not possible to ascertain that the definition of the homonym that conveys the meaning used in a particular document just read is included in the set of five alternatives. For example, a student might read a document that contained '…the weekly issue of Time magazine…' and thus the definition #2 for the *issue,* given in section 4, should be included as one of the five alternatives for the practice question. We refer to such correctly matched definitions as the 'true' definition for that (document, word) pair, in the following discussion. The version of REAP without WSD orders a word's dictionary definitions by their frequency of usage in CALD, and generates the five alternatives for a given word by choosing four distractor definitions for words of the same part of speech and by choosing the definition with highest usage frequency that has not yet been used in any of the previous questions for that student. This methodology has a better chance of including the 'true' definition in the five alternatives than a methodology based on random selection of definition. Nevertheless, it does not always guarantee inclusion of 'true' definition. These post-reading definition questions provide additional practice for the <word, sense> pair that the student was exposed to during the practice reading and thus reinforce the instruction. We claim that providing this additional exposure, where the student is actively involved in the process, promotes robust learning. Mismatched practice questions can potentially confuse the student and thus adversely affect student learning. Thus, studying the effects of this WSD-induced matching as measured by post-test performance is the most revealing comparative analysis.

Table 2 shows the data for this analysis. To make a fair comparison, we analyze only those <word, sense> pairs from the experimental group that would have been mismatched in the practice questions, even with the usage frequency methodology, had it not been for WSD. Thus, 45 <word, sense> pairs were found to have been well-matched in practice questions and texts because of WSD. The performance of the experimental group on these 45 <word, sense> pairs on the post-test is given in the Table 2. The columns split the data according to the information provided by the student during self-assessment. The rows group the data based on the post-test results. For the control group, only those <word, sense> pairs that did not get matched by chance for the practice questions are analyzed.

We use Generalized Estimating Equations (GEE) [18] to test our hypotheses. GEE takes into account the correlations between data-points that are not independently and

identically distributed. The use of GEE was warranted in this study because different data-points--corresponding to post-test results for particular words--were in some cases generated by the same student. Based on the data in Table 2 we perform the following hypothesis test using GEE.

H0: The true distribution of the post-test performance of the experimental group for the chosen words and the true distribution of the post-test performance of the control group is the same.

H1: The true distributions of post-test performance of the experimental group and the control group are not the same.

Analysis with GEE produces a *p*-value 0.001, which strongly supports rejecting the null hypothesis. The mean and the standard deviation for the experimental and control groups are ($M = 0.8$, $SD = 0.404$) and ($M = 0.5$, $SD = 0.505$), respectively. We also performed another analysis for testing the above hypotheses, however, this time two additional explanatory (independent) variables were also included, pre-test information and the MTELP score of the student. This analysis produced a *p*-value of 0.003 for the coefficient learned for the WSD status, which was the only significant coefficient in the model that was fit. Thus we can conclude that the treatment given to the experimental group had a statistically reliable and positive effect on their performance in the task of vocabulary learning.

Table 2. Data from the experimental and control groups of the user study.

	Experimental Group				**Control Group**		
	Known	**Unknown**			**Known**	**Unknown**	
Correct	28	8	36	**Correct**	16	6	22
Incorrect	7	2	9	**Incorrect**	16	6	22
	35	10	45		32	12	44

It is important to note that the pre-test required the students to indicate the words that they were familiar with, but did not ask about their familiarity with the <word, sense> pairs. As a result, although it appears from the data in Table 2 that most of the students were familiar with majority of the words included in this study, it is quite likely that a student who indicated being familiar with a word could only be familiar with only one of the meanings of the word. In fact, the second analysis above showed that the pre-test information could not explain students' performance on the post-test.

6 Conclusion

This work establishes that performing sense disambiguation for homonyms helps vocabulary learning in ESL students. It is demonstrated that the task of disambiguating homonyms can be automated by learning classifiers that can assign the appropriate sense to a homonym in a given context with high accuracy. A user study reveals that students equipped with WSD-enabled vocabulary tutor perform significantly better than students using vocabulary tutor without the WSD capabilities.

Acknowledgments
We thank Alan Juffs, Lois Wilson, Greg Mizera, Stacy Ranson and Chris Ortiz at the English Language Institute at the University of Pittsburgh for using REAP in the classroom. We also thank Dr. Howard Seltman and Dr. Ted Pedersen for their guidance and help. This work has been supported by the Barbara Lazarus Women@IT Fellowship, (in part) by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation, award number SBE-0354420 and Dept. of Education grant R305G03123. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsors.

References

1. Heilman, M., Collins-Thompson, K., Callan, J. & Eskenazi, M.: Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension. Proceedings of the Ninth International Conference on Spoken Language Processing. (2006)
2. Coxhead, A.: A New Academic Word List. TESOL, Quarterly, 34(2): 213-238. (2000).
3. Walter, E., editor.: Cambridge Advanced Learner's Dictionary, 2nd Edition. Cambridge University Press. (2005)
4. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, pp. 27--28 (2003)
5. Jorgenson, J. 1990. The psychological reality of word senses. Journal of Psycholinguistic Research 19:167-190.
6. Dolan, W. B.: "Word Sense Ambiguation: Clustering Related Senses," Proc. 15th Int'l. Conf. Computational Linguistics, ACL, Morristown, N.J., pp. 712-716. (1994).
7. Palmer, M., Dang, H. T., Fellbaum, C.: Making fine-grained and coarse-grained sense distinctions. Journal of Natural Language Engineering. (2005)
8. Ide, N., Jean, V.: Introduction to the special issue on word sense disambiguation: the state of the art. Computational Linguistics, 24(1). March, 1998. 1-40.
9. Mitchell, T. Machine Learning. The McGraw-Hill Companies, Inc. pp. 111-112, (1997).
10. Witten, I. H., Eibe, F.: Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, (2005).
11. McCallum, A., Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on "Learning for Text Categorization", (1998).
12. Vapnik, V. N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA. (1995).
13. Mohammad, S. and Pedersen, T.: Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation. Proceedings of the Conference on Computational Natural Language Learning, Boston, MA, (2004).
14. Larkey, L. S., & Croft, W. B.: Combining classifiers in text categorization. Proceedings of the Nineteenth Annual International ACM SIGIR Conference, pp. 289—297, (1996).
15. Schütze, H.: Automatic word sense discrimination. Computational Linguistics 24, 1, 97-123. (1998)
16. Pedersen, T., Bruce, R.: Distinguishing word senses in untagged text. Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, (1997).
17. Meara, P. and Buxton, B.: An alternative to multiple choice vocabulary tests. Language Testing, 4, 142–45. (1987).
18. Hardin, J. and Hilbe, J.: Generalized Estimating Equations. Chapman and Hall/CRC. (2003).