Reducing Annotation Effort on Unbalanced Corpus based on Cost Matrix

Wencan Luo, Diane Litman
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260, USA
{wencan,litman}@cs.pitt.edu

Joel Chan
Department of Psychology
University of Pittsburgh
Pittsburgh, PA 15260, USA
chozen86@gmail.com

Abstract

Annotated corpora play a significant role in many NLP applications. However, annotation by humans is time-consuming and costly. In this paper, a high recall predictor based on a cost-sensitive learner is proposed as a method to semi-automate the annotation of unbalanced classes. We demonstrate the effectiveness of our approach in the context of one form of unbalanced task: annotation of transcribed human-human dialogues for presence/absence of uncertainty. In two data sets, our cost-matrix based method of uncertainty annotation achieved high levels of recall while maintaining acceptable levels of accuracy. The method is able to reduce human annotation effort by about 80% without a significant loss in data quality, as demonstrated by an extrinsic evaluation showing that results originally achieved using manually-obtained uncertainty annotations can be replicated using semi-automatically obtained uncertainty annotations.

1 Introduction

Annotated corpora are crucial for the development of statistical-based NLP tools. However, the annotation of corpora is most commonly done by humans, which is time-consuming and costly. To obtain a higher quality annotated corpus, it is necessary to spend more time and money on data annotation. For this reason, one often has to accept some tradeoff between data quality and human effort.

A significant proportion of corpora are unbalanced, where the distribution of class categories are heavily skewed towards one or a few categories. Unbalanced corpora are common in a number of different tasks, such as emotion detection (Ang et al., 2002; Alm et al., 2005), sentiment classification (Li et al., 2012), polarity of opinion (Carvalho et al., 2011), uncertainty and correctness of student answers in tutoring dialogue systems (Forbes-Riley and Litman, 2011; Dzikovska et al., 2012), text classification (Forman, 2003), information extraction (Hoffmann et al., 2011), and so on[1].

In this paper, we present a semi-automated annotation method that can reduce annotation effort for the class of binary unbalanced corpora. Here is our proposed annotation scheme: the first step is to build a high-recall classifier with some initial annotated data with an acceptable accuracy via a cost-sensitive approach. The second step is to apply this classifier to the rest of the unlabeled data, where the data are then classified with positive or negative labels. The last step is to manually check every positive label and correct it if it is wrong.

To apply this method to work in practice, two research questions have to be addressed. The first one is how to get a high-recall classifier. High recall means only a low proportion of true positives are misclassified (false negatives). This property allows for only positive labels to be corrected by human annotators in the third step, so that annotation effort may be reduced. A related and separate research question concerns the overall quality of data when false negatives are not corrected: will a dataset annotated with this method produce the same results as a

[1]The unbalanced degrees - proportion of minority class category, of these corpora range from 3% to 24%.

Proceedings of the NAACL HLT 2013 Student Research Workshop, pages 8–15,
Atlanta, Georgia, 13 June 2013.

fully manually annotated version of the same dataset when analyzed for substantive research questions?

In this paper, we will answer the two research questions in the context of one form of binary unbalanced task[2]: annotation of transcribed human-human dialogue for presence/absence of uncertainty.

The contribution of this paper is twofold. First, an extrinsic evaluation demonstrates the utility of our approach, by showing that results originally achieved using manually-obtained uncertainty annotations can be replicated using semi-automatically obtained uncertainty annotations. Second, a high recall predictor based on a cost-sensitive learner is proposed as a method to semi-automate the annotation of unbalanced classes such as uncertainty.

2 Related Work

2.1 Reducing Annotation Effort

A number of semi-supervised learning methods have been proposed in the literature for reducing annotation effort, such as active learning (Cohn et al., 1994; Zhu and Hovy, 2007; Zhu et al., 2010), co-training (Blum and Mitchell, 1998) and self-training (Mihalcea, 2004). Active learning reduces annotation by carefully selecting more useful samples. Co-training relies on several conditional independent classifiers to tag new unlabeled data and self-training takes the advantage of full unlabeled data. These semi-supervised learning methods demonstrate that with a small proportion of annotated data, a classifier can achieve comparable performance with all annotated data. However, these approaches still need considerable annotation effort when a large corpus has to be annotated. In that case, all predicted labels have to be rechecked by humans manually. In addition, none of them take advantage of unbalanced data.

Another class of effort reduction techniques is pre-annotation, which uses supervised machine-learning systems to automatically assign labels to the whole data and subsequently lets human annotators correct them (Brants and Plaehn, 2000; Chiou et al., 2001; Xue et al., 2002; Ganchev et al., 2007; Chou et al., 2006; Rehbein et al., 2012).

Generally speaking, our annotation method belongs to the class of pre-annotation methods. However, our method improves pre-annotation for unbalanced data in two ways. Firstly, we lower the threshold for achieving a high recall classifier. Secondly, with pre-annotation, although people only perform a binary decision of whether the automatic classifier is either right or wrong, they have to go through all the unlabeled data one by one. In contrast, in our scheme, people go through only the positive predictions, which are much less than the whole unlabeled data, due to the unbalanced structure of the data. What's more, reducing the annotation effort is the goal of this paper but not building a high recall classifier such as Prabhakaran et al. (2012) and Ambati et al. (2010).

The approach proposed by Tetreault and Chodorow (2008) is similar to us. However, they assumed they had a high recall classifier but did not explicitly show how to build it. In addition, they did not provide extrinsic evaluation to see whether a corpus generated by pre-annotation is good enough to be used in real applications.

2.2 Uncertainty Prediction

Uncertainty is a lack of knowledge about internal state (Pon-Barry and Shieber, 2011). In this paper, we only focus on detection of uncertainty on text. Commonly used features are lexical features such as unigram (Forbes-Riley and Litman, 2011). Moreover, energy, dialogue features such as turn number, tutor goal, and metadata like gender are also considered by Forbes-Riley and Litman (2011). Uncertainty prediction is both substantively interesting (Chan et al., 2012; Forbes-Riley and Litman, 2009) and pragmatically expeditious for our purposes, due to its binary classification and typical unbalanced class structure.

CoNLL 2010 has launched a shared task to detect hedges and their scope in natural language text on two data sets: BioScope and Wikipedia (CoNLL, 2010). This first task to detect whether there is a hedge present or not present in a sentence is very similar to our uncertainty prediction task. 23 teams participated in the shared task with the best recall of 0.8772 on the BioScope, and 0.5528 on the Wikipedia. As we can see, uncertainty detection is not trivial and it can be hard to get a high recall classifier.

In this paper, we focus on lexical features for our

[2]This annotation scheme can also benefit other kinds of tasks.

purpose because lexical features are simple to extract and sufficient for our scheme. Even though other features may improve uncertainty prediction performance, with the goal of reducing annotation effort, such lexical features are shown to be good enough for our task.

3 The Corpora

We examine the following two data sets: the Mars Exploration Rover (MER) mission (Tollinger et al., 2006; Paletz and Schunn, 2011) and the student engineering team (Eng) dataset (Jang and Schunn, 2012). The MER scientists are evaluating data downloaded from the Rover, discussing their work process, and/or making plans for the Rovers. They come from a large team of about 100+ scientists/faculty, graduate students, and technicians. At any one time, conversations are between 2-10 people. The Eng teams are natural teams of college undergraduates working on their semester-long product design projects. The conversations involve 2-6 individuals. Audio and video are available for both data sets and transcripts are obtained with human annotators.

Our task is to annotate the transcribed human-human dialogues for presence/absence of uncertainty in each utterance. There are 12,331 transcribed utterances in the MER data set, and 44,199 transcribed utterances in the Eng data set. Both data sets are unbalanced: in the MER data, 1641 of all the 12,331 (13.3%) utterances are annotated as uncertain by trained human annotators; in the Eng data, only 1558 utterances are annotated, 221 of which are annotated as uncertain (14.2%). 96.5% of the utterances in the Eng data set have not been annotated yet, raising the need for an efficient annotated technique. Both data sets are annotated by two trained coders with high inter-rater agreement, at Cohen's kappa of 0.75 (Cohen, 1960). A sample dialogue snippet from the MER corpus is shown in Table 1. The last column indicates whether the utterance is labeled as uncertainty or not: '1' means uncertainty and '0' means certainty.

The MER data serves as the initial annotated set and a high recall classifier will be trained on it; the Eng data[3] serves as a simulated unlabeled data set to test the performance of our annotation scheme.

speaker	utterance	uncertainty?
S6	You can't see the forest through the trees.	0
S1	Yea, we never could see the [missing words]	1
S6	No we had to get above it	0
S4	We just went right through it	0
S6	Yea	0
S1	I still don't,	0
	I'm not quite sure	1

Table 1: Sample dialogue from the MER corpus

4 High Recall Classifier

4.1 Basic Classifier

The uncertainty prediction problem can be viewed as a binary classification problem. It involves two steps to build a high recall classifier for unbalanced data. The first step is to build up a simple classifier; the second step is to augment this classifier to favor high recall.

Aiming for a simple classifier with high recall, only some lexical words/phrases are used as features here. There are several resources for the words/phrases of uncertainty prediction. The main resource is a guideline book used by our annotators showing how to distinguish uncertainty utterance. It gives three different kinds of words/phrases, shown in Table 2 indicated by three superscripts '+', '-' and '*'. The words/phrases with '+' show some evidence of uncertainty; ones with '-' mean that they show no evidence of uncertainty; others with '*' may or may not show uncertainty. The second source is from existing literature. The words/phrases with '1' are from (Hiraishi et al., 2000) and ones with '2' are from (Holms, 1999).

For each word/phrase w, a binary feature is used to indicate whether the word/phrase w is in the utterance or not.

A Naive Bayes classifier is trained on the MER data using these features and tested on the Eng data. The performances of the model on the train set and test set are shown in Table 3. Both weighted and unweighted false positive (FP) Rate, Precision, Recall and F-Measure are reported. However, in later experiments, we will focus on only the positive class (the uncertainty class). A 0.689 recall means that 510 out of 1641 positive utterances are missed using this model.

[3]The Eng data in this paper denotes the annotated subset of the original Eng corpus.

as far as^{+}	i hope^{+}	somehow^{+}	it will^{-}	don't remember*	maybe*	tends to*	doubtful^{1}
as far as i know^{+}	i think^{+}	something^{+}	it wont^{-}	essentially*	most*	that can vary*	good chance^{1}
as far as we know^{+}	i thought^{+}	something like this^{+}	it would^{-}	fairly*	mostly*	typically*	improbable^{1}
believe^{+}	i wont^{+}	worried that^{+}	would it be^{-}	for the most part*	normally*	uh*	possible^{1}
could^{+}	im not sure^{+}	you cannot tell^{+}	about*	frequently*	pretty much*	um*	probable^{1}
guess^{+}	may^{+}	can^{-}	almost*	generally*	quite*	usually*	relatively^{1}
guessed^{+}	might^{+}	i am^{-}	any nonprecise amount*	hes*	should*	very*	roughly^{1}
guessing^{+}	not really^{+}	i can^{-}	basically*	hopefully*	sometimes*	virtually*	tossup^{1}
i believe^{+}	not sure^{+}	i will^{-}	believed*	i assumed that*	somewhat*	whatever*	unlikely^{1}
i cant really^{+}	possibly^{+}	i would^{-}	cannot remember*	it sounds as*	somewhere*	you know*	of course^{2}
i feel^{+}	probably^{+}	it can^{-}	can't remember*	kind of*	stuff*	almost certain^{1}	sort of^{2}
i guess^{+}	really^{+}	it is^{-}	do not remember*	likely*	tend to*	almost impossible^{1}	

Table 2: Words/phrases for uncertainty prediction.

Data Set	FP Rate	Precision	Recall	F-Measure	Class
MER	.311	.954	.989	.971	0
	.011	.908	.689	.784	1
	.271	.948	.949	.946	(Weighted)
Eng	.475	.926	.981	.952	0
	.019	.817	.525	.639	1
	.41	.91	.916	.803	(Weighted)

Table 3: Naive Bayes classifier performance on the MER (train set) and Eng (test set) with only the words/phrases

assume	I didn't know	more or less	some kind
couldn't	i don't even know	no idea	suppose
don't know	if	not clear	suspect
don't think	if it	or	think
don't understand	if we	perhaps	thought
doubt	if you	possibility	unclear
either	imagine	potential	what i understood
figured	kinda	presumably	wondering
i bet	kinds of	seem	
i can try	like	some	

Table 4: New words/phrases for uncertainty prediction

After error analysis, a few new words/phrases are added to the feature set, shown in Table 4. By supplementing the original feature set in this way, we reran the training yielding our final baseline, the performance on the training data (MER) and testing data (Eng) is shown in Table 5. This time, we compare different classifiers including Naive Bayes (NB), Decision Tree (DT) and Support Vector Machine (SVM). All of them are implemented using the open source platform Weka (Hall et al., 2009) with default parameters.

As we can see, test recall is worse than train recall.

Data Set	Method	TP	FP	Precision	Recall	F-Measure
MER	NB	.732	.016	.875	.732	.797
	DT	.831	.013	.908	.831	.868
	SVM	.811	.013	.905	.811	.855
Eng	NB	.679	.014	.888	.679	.769
	DT	.665	.021	.84	.665	.742
	SVM	.674	.022	.832	.674	.745

Table 5: Performance with original and new words/phrases as a feature set: train on the MER and test on the Eng data for class '1'. TP is true positive; FP is false positive

In addition, although DT and SVM perform better than NB on train data set, they have similar performance on the test set. Thus, the performance of the baseline is not unacceptable, but neither is it stellar. In advance, it is not hard to build such a model, since only simple features and classifiers are used here.

4.2 Augmenting the Classifier using a Cost Matrix

In our annotation framework, if the classifier achieves 100% recall, the annotated data will be perfect because all the wrong predictions can be corrected. That's the reason why we are seeking for a high recall classifier. A confusion matrix, is a common way to represent classifier performance. High recall is indexed by a low false negative (FN) rate; therefore, we aim to minimize FNs to achieve high recall.

Following this idea, we employ a cost-sensitive model, where the cost of FN is more than false positive (FP).

Following the same notation, we represent our cost-sensitive classifier as a cost matrix. In our cost matrix, classifying an actual class '1' as '1' costs C_{tp}, an actual class '0' as '1' costs C_{fp}, an actual class '1' to '0' costs C_{fn}, and '0' to '0' costs C_{tn}. To achieve a high recall, C_{fn} should be more than C_{fp}.

We can easily achieve 100% recall by classifying all samples to '1', but this would defeat our goal of reducing human annotation effort, since all utterance uncertainty predictions would need to be manually corrected. Thus, at the same time of a high recall, we should also balance the total ratio of TP and FP.

In our experiment, C_{tp} and C_{tn} are set to 0 since they are perfectly correct. Additionally, $C_{fp} = 1$ all the time and C_{fn} changes with different scales. FPs

C_{fn}	FP Rate	Precision	Recall	F-Measure	$(TP+FP)/N$
1	.022	.831	.67	.742	.114
2	.024	.825	.683	.748	.117
3	.037	.771	.747	.759	.138
5	.052	.726	.828	.774	.162
10	.071	.674	.887	.766	.187
15	.091	.622	.91	.739	.207
20	.091	.622	.91	.739	.207

Table 6: Test performance with cost matrix

mean wrong predictions, but we can correct them during the second pass to check them. However, we cannot correct FNs without going through the whole data set, so they are a more egregious detriment to the quality of the annotated data. During the experiment, C_{fn} varies from 1 to 20. With increases in C_{fn}, the cost of FN increases compared to FP.

The cost-sensitive classifier is relying on Weka with reweighting training instances. In this task, SVM performed better than NB and DT. Only SVM results are included here due to space constraint. The test results are shown in Table 6[4]. The last column in the two tables is the total proportion of positive predictions ($FP + TP$). This value indicates the total amount of data that humans have to check in the second pass to verify whether positive predictions are correct. To reduce human annotation effort, we would like this value to be as low as possible.

As shown in Table 6, with the increase of C_{fn}, the recall increases; however, the proportion of positive predictions also increases. Therefore, it is a tradeoff to achieve a high recall and a low ratio of TP and FP.

For the test set, the recall increases with larger C_{fn}, even with a small increase of C_{fn} from 1 to 3. Remarkably, the classifier gives us a high recall while keeping the proportion of positive predictions at an acceptably low level. When $C_{fn} = 20$ for the test set, only 20.7% of the data need to be manually checked by humans, and less than 10% uncertain utterances (19 out of 221 for the Eng data) are missed.

Now, we have achieved a high recall classifier with an acceptable ratio of positive predictions.

5 Extrinsic Evaluation of Semi-Automated Annotation

Even with a high recall classifier, some of the true positive data are labeled incorrectly in the final annotated corpus. In addition, it also changes the distribution of class labels.

To test whether it hurts the overall data quality, we performed an analysis, which demonstrates that this annotation scheme is sufficient to produce quality data. We attempted to replicate an analysis on the Eng data set, which examines the use of analogy, a cognitive strategy where a source and target knowledge structure are compared in terms of structural correspondences as a strategy for solving problems under uncertainty. The analysis we attempt to replicate here focuses on examining how uncertainty levels change relative to baseline before, during, and after the use of analogies.

The overall Eng transcripts were segmented into one of 5 block types: 1) pre-analogy (Lag -1) blocks, 10 utterances just prior to an analogy episode, 2) during-analogy (Lag 0) blocks, utterances from the beginning to end of an analogy episode, 3) post-analogy (Lag 1) blocks, 10 utterances immediately following an analogy episode, 4) post-post-analogy (Lag 2) blocks, 10 utterances immediately following post-analogy utterances, and 5) baseline blocks, each block of 10 utterances at least 25 utterances away from the other block types. The measure of uncertainty in each block was the proportion of uncertain utterances. The sampling strategy for the baseline blocks was designed to provide an estimate of uncertainty levels when the speakers were engaged in pre-analogy, during-analogy, or post-analogy conversation, with the logic being that a certain amount of lag or spillover of uncertainty was assumed to take place surrounding analogy episodes.

Figure 1 shows the relationship of block type to mean levels of uncertainty, comparing the pattern with human vs. classifier-supported uncertainty labels. The classifier-generated labels were first preprocessed such that all FPs were removed, but FNs remain. This re-analysis comparison thus provides a test of whether the recall rate is high enough that known statistical effects are not substantially altered or removed. To examine how different settings of C_{fn} might impact overall performance, we used labels (corrected for false positives) for 4 different levels of C_{fn} (1, 5, 10, 20) from the Table 6.

In the Eng data analyses, the main findings were that analogy was triggered by local spikes in uncertainty levels (Lag -1 $>$ baseline), replicating re-

[4]Only $C_{fn} = 1, 2, 3, 5, 10, 15, 20$ are reported here due to page limits

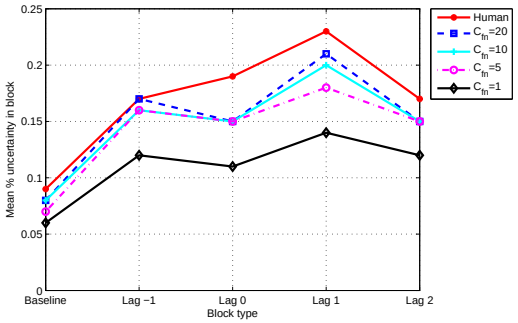

Figure 1: Mean % uncertainty by block type and label source (Eng data set)

Table 7: Standardized mean difference (Cohen's d) from baseline by block type and label source (the Eng data set) (Note: '*' denotes $p < .05$, '**' denotes $p < .01$)

	Block type			
	Lag -1	Lag 0	Lag 1	Lag 2
Human	0.54*	0.4	0.79**	0.46*
$C_{fn} = 20$	0.57*	0.3	0.78**	0.44
$C_{fn} = 10$	0.58**	0.32	0.73**	0.47*
$C_{fn} = 5$	0.57*	0.34	0.66**	0.48*
$C_{fn} = 1$	0.42	0.25	0.54*	0.40

sults from prior work with the MER dataset (Chan et al., 2012); in contrast to the findings in MER, uncertainty did not reduce to baseline levels following analogy (Lags 1 and 2 > baseline). Figure 1 plots the relationship of block type to mean levels of uncertainty in this data set, comparing the pattern with human vs. classifier-generated uncertainty labels. Table 7 shows the standardized mean difference (Cohen's d) (Cohen, 1988) from baseline by block type and label source. The pattern of effects (Lag -1 > baseline, Lags 1 and 2 > baseline) remains substantially unchanged with the exception of the Lag 2 vs. baseline comparison falling short of statistical significance (although note that the standardized mean difference remains very similar) for C_{fn} ranging from 20 to 5, although we can observe a noticeable attenuation of effect sizes from C_{fn} of 5 and below, and a loss of statistical significance for the main effect of uncertainty being significantly higher than baseline for Lag -1 blocks when C_{fn} = 1.

The re-analysis clearly demonstrates that the recall rate of the classifier is sufficient to not substantially alter or miss known statistical effects. We can reasonably extrapolate that using this classifier for uncertainty annotation in other datasets should be satisfactory.

6 Conclusion and Discussion

In this paper, a simple high recall classifier is proposed based on a cost matrix to semi-automate the annotation of corpora with unbalanced classes. This classifier maintains a good balance between high recall and high FP and NP ratio. In this way, humans can employ this classifier to annotate new data with significantly reduced effort (approximately 80% less effort, depending on the degree of imbalance in the data). Although the classifier does introduce some misclassified samples to the final annotation, an extrinsic evaluation demonstrates that the recall rate is high enough and the performance does not sacrifice data quality.

Like other semi-supervised or supervised methods for supporting annotation, our annotation scheme has some limitations that should be noted. Firstly, an initial annotated data set is needed to derive a good performance classifier and the amount of annotated data is dependent on the specific task[5]. Secondly, the features and machine learning algorithms used in semi-supervised annotation are also domain specific. At the same time, there are some unique challenges and opportunities that can be further investigated for our annotation scheme on unbalanced data. For example, even though the cost matrix method can achieve a high recall for binary classification problem, whether it can be generalized to other tasks (e.g., multi-class classification tasks) is an unanswered question. Another open question is how the degree of unbalance between classes in the corpora affects overall annotation quality. We suggest that if the data is not unbalanced, the total amount of effort that can be reduced will be lower.

Acknowledgments

The collection of the engineering data was supported by NSF grants SBE-0738071, SBE-0823628, and SBE-0830210. Analogy analysis was supported by NSF grant SBE-1064083.

[5]For a new task, a new feature set is usually derived.

References

Cecilia Ovesdotter Alm, Dan Roth and Richard Sproat. 2005. *Emotions from text: Machine learning for text-based emotion prediction*. In Proceedings of HLT/EMNLP 2005.

Bharat Ram Ambati, Mridul Gupta, Samar Husain and Dipti Misra Sharma. 2010. *A high recall error identification tool for Hindi treebank validation*. In Proceedings of The 7th International Conference on Language Resources and Evaluation (LREC), Valleta, Malta.

Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg and Andreas Stolcke. 2002. *Prosody-based automatic detection of annoyance and frustration in human-computer Dialog*. In INTERSPEECH-02.

Avrim Blum and Tom Mitchell. 1998. *Combining labeled and unlabeled data with co-training*. In Proceedings of the eleventh annual conference on Computational learning theory, p.92-100, July 24-26, Madison, Wisconsin, United States

Thorsten Brants and Oliver Plaehn. 2000. *Interactive corpus annotation*. In Proceedings of LREC-2000.

Paula Carvalho, Luís Sarmento, Jorge Teixeira and Mário J. Silva. 2011. *Liars and saviors in a sentiment annotated corpus of comments to political debates*. In Proceedings of the Association for Computational Linguistics (ACL 2011), Portland, OR.

Joel Chan, Susannah B. F. Paletz and Christian D. Schunn. 2012. *Analogy as a strategy for supporting complex problem solving under uncertainty*. Memory & Cognition, 40, 1352-1365.

Fu-Dong Chiou, David Chiang and Martha Palmer. 2001. *Facilitating treebank annotation using a statistical parser*. In HLT'01. ACL.

Wen-Chi Chou, Richard Tzong-Han Tsai, Ying-Shan Su, Wei Ku1, Ting-Yi Sung and Wen-Lian Hsu. 2006. *A semi-automatic method for annotating a biomedical proposition bank*. In Proceedings of FLAC-2006.

David Cohn, Richard Ladner and Alex Waibel. 1994. *Improving generalization with active learning*. Machine Learning, 15 (2), 201-221.

Jacob Cohen. 1960. *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement, 20, 37-46.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences (2nd ed.)*. Lawrence Erlbaum.

CoNLL-2010 Shared Task. 2010. In Fourteenth Conference on Computational Natural Language Learning, Proceedings of the Shared Task.

Myroslava Dzikovska, Peter Bell, Amy Isard and Johanna D. Moore. 2012. *Evaluating language understanding accuracy with respect to objective outcomes in a dialogue system*. EACL 2012: 471-481.

Kate Forbes-Riley and Diane Litman. 2009. *Adapting to student uncertainty improves tutoring dialogues*. In Proceedings 14th International Conference on Artificial Intelligence in Education (AIED2009), pp. 33-40.

Kate Forbes-Riley and Diane Litman. 2011. *Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor*. Speech Communication, v53, pp. 1115-1136.

George Forman 2003. *An Extensive empirical study of feature selection metrics for text classification*. Journal of Machine Learning Research, 3, 1289-1305.

Kuzman Ganchev, Fernando Pereira, Mark Mandel, Steven Carroll and Peter White. 2007. *Semi-automated named entity annotation*. In Proceedings of the linguistic annotation workshop, pp. 53-56

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H.Witten. 2009. *The WEKA data mining software: An update*. SIGKDD Explorations, 11(1).

Taka Hiraishi, Buruhani Nyenzi, Jim Penman and Semere Habetsion. 2000. *Quantifying uncertainties in practice*. In Revised 1996 IPCC guidelines for national greenhouse gas inventories.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer and Daniel S. Weld. 2011. *Knowledge-based weak supervision for information extraction of overlapping relations*. In ACL.

Janet Holmes. 1999. *Women, men, and politeness*. London, SAGE publications, pp:86-96

Jooyoung Jang and Christian Schunn. 2012. *Physical design tools support and hinder innovative engineering design*. Journal of Mechanical Design, vol. 134, no. 4, pp. 041001-1-041001-9.

Shoushan Li, Shengfeng Ju, Guodong Zhou and Xiaojun Li. 2012. *Active learning for imbalanced sentiment classification*. EMNLP-CoNLL 2012: 139-148

Rada Mihalcea. 2004. *Co-training and self-training for word sense disambiguation*. In Proceedings of the 8th Conference on Computational Natural Language Learning (CoNLL, Boston, MA). 33-40.

Susannah B. F. Paletz and Christian D. Schunn. 2011. *Assessing group-level participation in fluid teams: Testing a new metric*. Behav Res 43:522-536.

Heather Pon-Barry and Stuart M. Shieber 2011. *Recognizing uncertainty in speech*. EURASIP Journal on Advances in Signal Processing.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow and Benjamin Van Durme 2012 *Statistical modality tagging from rule-based annotations and crowdsourcing*. In Proceedings of ACL Workshop on Extra-propositional aspects of meaning in computational linguistics (ExProM).

Ines Rehbein, Josef Ruppenhofer and Caroline Sporleder. 2012. *Is it worth the effort? Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation*. Language Resources and Evaluation, Vol.46, No.1. pp. 1-23

Joel R. Tetreault and Martin Chodorow. *Native judgments of non-native usage: experiments in preposition error detection*. In Proceedings of the Workshop on Human Judgements in Computational Linguistics, p.24-32, Manchester, United Kingdom.

Irene V. Tollinger, Christian D. Schunn and Alonso H. Vera. 2006. *What changes when a large team becomes more expert? Analyses of speedup in the Mars Exploration Rovers science planning process*. In Proceedings of the 28th Annual Conference of the Cognitive Science Society (pp. 840-845). Mahwah, NJ: Erlbaum.

Nianwen Xue, Fu-Dong Chiou and Martha Palmer. 2002. *Building a large-scale annotated chinese corpus*. In Proceedings of the 19th international conference on Computational linguistics. ACL.

Jingbo Zhu and Eduard Hovy. 2007. *Active learning for word sense disambiguation with methods for addressing the class imbalance problem*. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 783-790.

Jingbo Zhu, Huizhen Wang, Eduard H. Hovy and Matthew Y. Ma. 2010. *Confidence-based stopping criteria for active learning for data annotation*. ACM Transactions on Speech and Language Processing, 6, 124.