# Latent Features in Automatic Tense Translation between Chinese and English

**Yang Ye**[†]**, Victoria Li Fossum**[§]**, Steven Abney** [† §]
[†] Department of Linguistics
[§] Department of Electrical Engineering and Computer Science
University of Michigan

**Abstract**

On the task of determining the tense to use when translating a Chinese verb into English, current systems do not perform as well as human translators. The main focus of the present paper is to identify features that human translators use, but which are not currently automatically extractable. The goal is twofold: to test a particular hypothesis about what additional information human translators might be using, and as a pilot to determine where to focus effort on developing automatic extraction methods for features that are somewhat beyond the reach of current feature extraction. The paper shows that incorporating several latent features into the tense classifier boosts the tense classifier's performance, and a tense classifier using only the latent features outperforms one using only the surface features. Our findings confirm the utility of the latent features in automatic tense classification, explaining the gap between automatic classification systems and the human brain.

## 1 Introduction

Language speakers make two types of distinctions about temporal relations: the first type of relation is based on precedence between events and can be expanded into a finer grained taxonomy as proposed by (Allen, 1981). The second type of relation is based on the relative positioning between the following three time parameters proposed by (Reichenbach, 1947): speech time (S), event time (E) and reference time (R). In the past couple of decades, the NLP community has seen an emergent interest in the first type of temporal relation. In the cross-lingual context, while the first type of relationship can be easily projected across a language pair, the second type of relationship is often hard to be projected across a language pair. In contrast to this challenge, cross-lingual temporal reference distinction has been poorly explored.

Languages vary in the granularity of their tense and aspect representations; some have finer-grained tenses or aspects than others. Tense generation and tense understanding in natural language texts are highly dynamic and context-dependent processes, since any previously established time point or interval, whether explicitly mentioned in the context or not, could potentially serve as the reference time for the event in question. (Bruce, 1972) captures this nature of temporal reference organization in discourse through a multiple temporal reference model. He defines a set ($S_1$, $S_2$, ..., $S_n$) that is an element of tense. $S_1$ corresponds to the speech time, $S_n$ is the event time, and ($S_i$, i=2, ..., n-1) stand for a sequence of time references from which the reference time of a particular event could come. Given the elusive nature of reference time shift, it is extremely hard to model the reference time point directly in temporal information processing. The above reasons motivate classifying temporal reference distinction automatically, using machine learning algorithms such as Conditional Random Fields (CRFs).

Many researchers in Natural Language Processing seem to believe that an automatic system does not have to follow the mechanism of human brain in order to optimize its performance, for example, the feature space for an automatic classification system does not have to replicate the knowledge sources that human beings utilize. There has been very little research that pursues to testify this faith.

The current work attempts to identify which features are most important for tense generation in Chinese to English translation scenario, which can point to direction of future research effort for automatic tense translation between Chinese and English.



*Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 48–55,
Sydney, July 2006. 

The remaining part of the paper is organized as follows: Section 2 summarizes the significant related works in temporal information annotation and points out how this study relates to yet differs from them. Section 3 formally defines the problem, tense taxonomy and introduces the data. Section 4 discusses the feature space and proposes the latent features for the tense classification task. Section 5 presents the classification experiments in Conditional Random Fields as well as Classification Tree and reports the evaluation results. Section 6 concludes the paper and section 7 points out directions for future research.

## 2 Related Work

There is an extensive literature on temporal information processing. (Mani, et al., 2005) provides a survey of works in this area. Here, we highlight several works that are closely related to Chinese temporal information processing. (Li, 2001) describes a model of mining and organizing temporal relations embedded in Chinese sentences, in which a set of heuristic rules are developed to map linguistic patterns to temporal relations based on Allen's thirteen relations. Their work shows promising results via combining machine learning techniques and linguistic features for successful temporal relation classification, but their work is concerned with another type of temporal relationship, namely, the precedence-based temporal relation between a pair of events explicitly mentioned in text.

A significant work worth mentioning is (Olsen et. al. 2001)'s paper, where the authors examine the determination of tense for English verbs in Chinese-to-English translation. In addition to the surface features such as the presence of aspect markers and certain adverbials, their work makes use of the telicity information encoded in the lexicons through the use of Lexical Conceptual Structures (LCS). Based on the dichotomy of grammatical aspect and lexical aspect, they propose that past tense corresponds to the telic (either inherently or derived) LCS. They propose a heuristic algorithm in which grammatical aspect markings supersede the LCS, and in the absence of grammatical aspect marking, verbs that have telic LCS are translated into past tense and present tense otherwise. They report a significant performance improvement in tense resolution from adding a verb telicity feature. They also achieve better performance than the baseline system using the telicity feature alone. This work, while alerting researchers to the importance of lexical aspectual feature in determination of tense for English verbs in Chinese-to-English machine translation, is subject to the risk of adopting a one-to-one mapping between grammatical aspect markings and tenses hence oversimplifies the temporal reference problem in Chinese text. Additionally, their binary tense taxonomy is too coarse for the rich temporal reference system in Chinese.

(Ye, et al. 2005) reported a tense tagging case study of training Conditional Random Fields on a set of shallow surface features. The low inter-annotator agreement rate reported in the paper illustrates the difficulty of tense tagging. Nevertheless, the corpora size utilized is too small with only 52 news articles and none of the latent features was explored, so the evaluation result reported in the paper leaves room for improvement.

## 3 Problem Definition

### 3.1 Problem Formulation

The problem we are interested in can be formalized as a standard classification or labeling problem, in which we try to learn a classifier

$$C : V \rightarrow T \tag{1}$$

where V is a set of verbs (each described by a feature vector), and T is the set of possible tense tags.

Tense and aspect are morphologically merged in English and coarsely defined, there can be twelve combinations of the simple tripartite tenses (present, past and future) with the progressive and perfect grammatical aspects. For our classification experiments, in order to combat sparseness, we ignore the aspects and only deal with the three simple tenses: present, past and future.

### 3.2 Data

We use 152 pairs of parallel Chinese-English articles from LDC release. The Chinese articles come from two news sources: Xinhua News Service and Zaobao News Service, consisting of 59882 Chinese characters in total with roughly 350 characters per article. The English parallel articles are from Multiple-Translation Chinese (MTC) Corpus from LDC with catalog number LDC2002T01. We chose to use the best human translation out of 9 translation teams as our gold-standard parallel English data. The verb tenses are obtained through manual alignment between the Chinese source articles and the English translations. In order to avoid the noise brought by errors and be focused on the central question we try to answer in the paper, we did not use automatic tools such as GIZA++ to obtain the verb alignments, which typically comes with significant amount of errors. We ignore Chinese verbs that are not translated into English as verbs because of "nominalization" (by which verbal expressions in Chinese are translated into nominal phrases in English). This exclusion is based on the rationale that another choice of syntactic structure might retain the verbal status in the target English sentence, but the tense of those potential English verbs would be left to the joint decision of a set of disparate features. Those tenses are unknown in our training data. This preprocessing yields us a total of 2500 verb tokens in our data set.

## 4 Feature Space

### 4.1 Surface Features

There are many heterogeneous features that contribute to the process of tense generation for Chinese verbs in the cross-lingual situation. Tenses in English, while manifesting a distinction in temporal reference, do not always reflect this distinction at the semantic level, as is shown in the sentence "I will leave when he comes." (Hornstein, 1990) accounts for this phenomenon by proposing the Constraints on Derived Tense Structures. Therefore, the feature space we use includes the features that contribute to the semantic level temporal reference construction as well as those contributing to tense generation from that semantic level. The following is a list of the surface features that are directly extractable from the training data:

1. Feature 1: Whether the verb is in quoted speech or not.
2. Feature 2: The syntactic structure in which the current verb is embedded. Possible structures include sentential complements, relative clauses, adverbial clauses, appositive clauses, and null embedding structure.
3. Feature 3: Which of the following signal adverbs occur between the current verb and the previous verb: yi3jing1(already), ceng2jing1(once), jiang1(future tense marker), zheng4zai4(progressive aspect marker), yi4zhi2(have always been).
4. Feature 4: Which of the following aspect markers occur between the current verb and the subsequent verb: le0, zhe0, guo4.
5. Feature 5: The distance in characters between the current verb and the previously tagged verb (We descretize the continuous distance into three ranges: $0 < distance < 5$, $5 \leq distance < 10$, or $10 \leq distance < \infty$).
6. Feature 6: Whether the current verb is in the same clause as the previous verb.

Feature 1 and feature 2 are used to capture the discrepancy between semantic tense and syntactic tense. Feature 3 and feature 4 are clues or triggers of certain aspectual properties of the verbs. Feature 5 and feature 6 try to capture the dependency between tenses of adjacent verbs.

### 4.2 Latent Features

The bottleneck in Artificial Intelligence is the unbalanced knowledge sources shared by human beings and a computer system. Only a subset of the knowledge sources used by human beings can be formalized, extracted and fed into a computer system. The rest are less accessible and are very hard to be shared with a computer system. Despite their importance in human language processing, latent features have received little attention in feature space exploration in most NLP tasks because they are impractical to extract. Although there have not yet been rigorous psycholinguistic studies demonstrating the extent to which the above knowledge types are used in human temporal relation processing, we hypothesize that they are very significant in assisting human's temporal relation decision. Nevertheless, a quantitative assessment of the utility of the latent features in NLP tasks has yet to be explored. (Olsen, et al., 2001) illustrates the value of latent features by showing how the telicity feature alone can help with tense resolution in Chinese to English machine translation. Given the prevalence of latent features in human language processing, in order to emulate human beings performance of the disambiguation, it is crucial to experiment with the latent features in automatic tense classification.

(Pustejovsky, 2004) discusses the four basic problems in event-temporal identification:

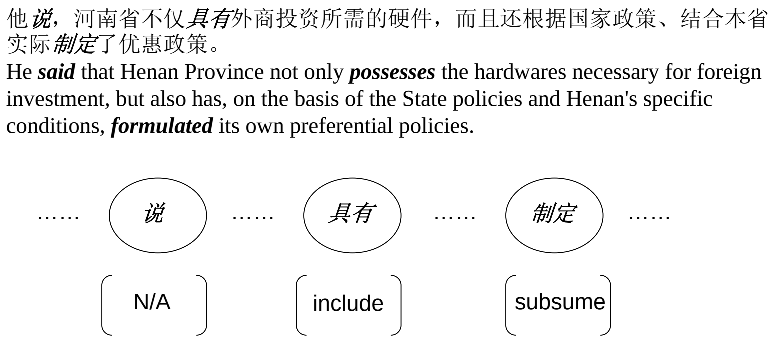


Figure 1: Temporal Relations between Adjacent Events

1. Time-stamping of events (identifying an event and anchoring it in time)

2. Ordering events with respect to one another

3. Reasoning with contextually under-specified temporal expressions

4. Reasoning about the persistence of events (how long does an event or the outcome of an event last?)

While time-stamping of the events and reasoning with contextually under-specified temporal expressions might be too informative to be features in tense classification, information concerning orderings between events and persistence of events are relatively easier to be encoded as features in a tense classification task. Therefore, we experiment with these two latent knowledge sources, both of which are heavily utilized by human beings in tense resolution.

### 4.3 Telicity and Punctuality Features

Following (Vendler, 1947), temporal information encoded in verbs is largely captured by some innate properties of verbs, of which telicity and punctuality are two very important ones. Telicity specifies a verb's ability to be bound in a certain time span, while punctuality specifies whether or not a verb is associated with a point event in time. Telicity and punctuality prepare verbs to be assigned different tenses when they enter the context in the discourse. While it is true that isolated verbs are typically associated with certain telicity and punctuality features, such features are contextually volatile. In reaction to the volatility exhibited in verb telicity and punctuality features, we propose that verb telicity and punctuality features should be evaluated only at the clausal or sentential level for the tense classification task. We manually obtained these two features for both the English and the Chinese verbs. All verbs in our data set were manually tagged as “telic” or “atelic”, and “punctual” or “apunctual”, according to context.

### 4.4 Temporal Ordering Feature

(Allen, 1981) defines thirteen relations that could possibly hold between any pair of situations. We experiment with six temporal relations which we think represent the most typical temporal relationships between two events. We did not adopt all of the thirteen temporal relationships proposed by Allen for the reason that some of them would require excessive deliberation from the annotators and hard to implement. The six relationships we explore are as follows:

1. event A precedes event B

2. event A succeeds event B

3. event A includes event B

4. event A subsumes event B

5. event A overlaps with event B

6. no temporal relations between event A and event B

For each Chinese verb in the source Chinese texts, we annotate the temporal relation between the verb and the previously tagged verb as belonging to one of the above classes. The annotation of the temporal relation classes mimics a deeper semantic analysis of the Chinese source text. Figure 1 illustrates a sentence in which each verb is tagged by the temporal relation class that holds between it and the previous verb.

## 5 Experiments and Evaluation

### 5.1 CRF learning algorithms

Conditional Random Fields (CRFs) are a formalism well-suited for learning and prediction on sequential data in many NLP tasks. It is a probabilistic framework proposed by (Lafferty et al., 2001) for labeling and segmenting structured data, such as sequences, trees and lattices. The conditional nature of CRFs relaxes the independence assumptions required by traditional Hidden Markov Models (HMMs). This is because the conditional model makes it unnecessary to explicitly represent and model the dependencies among the input variables, thus making it feasible to use interacting and global features from the input. CRFs also avoid the label bias problem exhibited by maximum entropy Markov models (MEMMs) and other conditional Markov models based on directed graphical models. CRFs have been shown to perform well on a number of NLP problems such as shallow parsing (Sha and Pereira, 2003), table extraction (Pinto et al., 2003), and named entity recognition (McCallum and Li, 2003). For our experiments, we use the MALLET implementation of CRF's (McCallum, 2002).

### 5.2 Experiments

#### 5.2.1 Human Inter-Annotator Agreement

All supervised learning algorithms require a certain amount of training data, and the reliability of the computational solutions is intricately tied to the accuracy of the annotated data. Human annotations typically suffer from errors, subjectivity, and the expertise effect. Therefore, researchers use consistency checking to validate human annotation experiments. The Kappa Statistic (Cohen, 1960) is a standard measurement of inter-annotator agreement for categorical data annotation. The Kappa score is defined by the following formula, where P(A) is the observed agreement rate from multiple annotators and P(E) is the expected rate of agreement due to pure chance:

$$k = \frac{P(A) - P(E)}{1 - P(E)} \tag{2}$$

Since tense annotation requires disambiguating grammatical meaning, which is more abstract than lexical meaning, one would expect the challenge posed by human annotators in a tense annotation experiment to be even greater than for word sense disambiguation. Nevertheless, the tense annotation experiment carried as a precursor to our tense classification task showed a kappa Statistic of 0.723 on the full taxonomy, with an observed agreement of 0.798. In those experiments, we asked three bilingual English native speakers who are fluent in Chinese to annotate the English verb tenses for the first 25 Chinese and English parallel news articles from our training data.

We could also obtain a measurement of reliability by taking one annotator as the gold standard at one time, then averaging over the precisions of the different annotators across different gold standards. While it is true that numerically, precision would be higher than Kappa score and seems to be inflating Kappa score, we argue that the difference between Kappa score and precision is not limited to one measure being more aggressive than the other. Rather, the policies of these two measurements are different. The Kappa score cares purely about agreement without any consideration of trueness or falseness, while the procedure we described above gives equal weight to each annotator being the gold standard, and therefore considers both agreement and truthness of the annotation. The advantage of the precision-based agreement measurement is that it makes comparison of the system performance accuracy to the human performance accuracy more direct. The precision under such a scheme for the three annotators is 80% on the full tense taxonomy.

#### 5.2.2 CRF Learning Experiments

We train a tense classifier on our data set in two stages: first on the surface features, and then on the combined space of both surface features (discussed in 4.1) and latent features (discussed in 4.2-4.4). It is conceivable that the granularity of sequences may matter in learning from data with sequential relationship, and in the context of verb tense tagging, it naturally maps to the granularity of discourse. (Ye, et al., 2005) shows that there is no significant difference between sentence-level sequences and paragraph-level sequences. Therefore, we experiment with only sentence-level sequences.

#### 5.2.3 Classification Tree Learning Experiments

To verify the stability of the utility of the latent features, we also experiment with classification tree learning on the same features space as

| Tense | Precision | Recall | F |
|---|---|---|---|
| Present tense | 0.662 | 0.661 | 0.627 |
| Past tense | 0.882 | 0.915 | 0.896 |
| Future tense | 0.758 | 0.487 | 0.572 |

Table 1: Evaluation Results for CRFs Classifier in Precision, Recall and F Using All Features

| | Surface Features | Latent Features | Surface and Latent Features |
|---|---|---|---|
| Accuracy for Training Data | 79.3% | 82.9% | 85.9% |

Table 2: Apparent Accuracy for the Training Data of the Classification Tree Classifiers

discussed above. Classification Trees are used to predict membership of cases or objects in the classes of a categorical dependent variable from their measurements on one or more predictor variables. The main idea of Classification Tree is to do a recursive partitioning of the variable space to achieve good separation of the classes in the training dataset. We use the Recursive Partitioning and Regression Trees(Rpart) package provided by R statistical computing software for the implementation of classification trees. In order to avoid over-fitting, we prune the tree by setting the minimum number of objects in a node to attempt a split and the minimum number of objects in any terminal node to be 10 and 3 respectively. In the constructed classification tree when we use all features including both surface and latent features, the top split at the root node in the tree is based on telicity feature of the English verb, indicating the importance of telicity feature for English verb among all of the features.

### 5.3 Evaluation Results

All results are obtained by 5-fold cross validation. The classifier's performance is evaluated against the tenses from the best-ranked human-generated English translation. To evaluate the performance of the CRFs tense classifier, we compute the precision, recall, general accuracy and F, which are defined as follow.

$$Accuracy = \frac{n_{prediction}}{N_{prediction}} \tag{3}$$

$$Recall = \frac{n_{hit}}{S} \tag{4}$$

$$Precision = \frac{n_{hit}}{N_{hit}} \tag{5}$$

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

where

1. $N_{prediction}$: Total number of predictions;
2. $n_{prediction}$: Number of correct predictions;
3. $N_{hit}$: Total number of hits;
4. $n_{hit}$: Number of correct hits;
5. $S$: Size of perfect hitlist;

From Table 1, we see that past tense, which occurs most frequently in the training data, has the highest precision, recall and F. Future tense, which occurs least frequently, has the lowest F. Precision and recall do not show clear pattern across different tense classes.

Table 2 presents the apparent classification accuracies for the training data, we see that latent features still outperform the surface features. Table 3 summarizes the general accuracies of the tense classification systems for CRFs and Classification Trees. The CRFs classifier and the Classification Tree classifier demonstrate similar scales of improvement from surface features, latent features to both surface and latent features.

| Methodology | Surface Features | Latent Features | Surface and Latent Features |
|---|---|---|---|
| CRFs | 75.8% | 80% | 83.4% |
| Classification Tree | 74.1% | 81% | 84.5% |

Table 3: Evaluations in General Accuracy

### 5.4 Baseline Systems

To better evaluate our tense classifiers, we provide two baseline systems here. The first baseline system is the tense resolution from the best ranked machine translation system's translation results in the MTC corpus mentioned above. When evaluated against the reference tense tags from the best ranked human translation team, the best MT system yields a accuracy of 47%. The second baseline system is a naive system that assigns the most frequent tense in the training data set, which in our case is past tense, to all verbs in the test data set. Given the fact that we are deadling with newswire data, this baseline system yields a high baseline system with an accuracy of 69.5%.

## 6 Discussion and Conclusions

To the best of our knowledge, the current paper is the first work investigating the utility of latent features in the task of machine-learning based automatic tense classification. We significantly outperform the two baseline systems as well as the automatic tense classifier performance reported by (Ye, et al., 2005) by 15% in general accuracy. A crucial finding of our experiments is that utility of only three latent features, i.e. verb telicity, verb punctuality and temporal ordering between adjacent events, outperforms that of all the surface linguistic features we discussed earlier in the paper. While one might think that the lack of existing techonology of latent feature extraction would discount research effort on latent features' utilities, we believe that such efforts guide the research community to determine where to focus effort on developing automatic extraction methods for features that are beyond the reach of current technologies. Such research effort will also help to shed light on the enigmatic research question of whether automatic NLP systems should take effort to make use of the features employed by human beings to optimize the system performance and shorten the gap between the system and human brain. The results of the current paper point to the fact that bottleneck of cross-linguistic tense classification is acquisition and modeling of the more latent linguistic knowledge. To our surprise, CRF tense classifier performance is consistently tied with classification tree tense classifier performance in all of our experiments. One might expect that CRFs would accurately capture sequential dependencies among verbs. Reflecting upon the similar evaluation results of the CRFs classifier and the Classification Tree classifier, it is unlikely for this to be due to the over-fitting of the Classification Tree because of the pruning we did to the Classification Trees. Therefore, we speculate that the dependencies between the tense tags of verbs in the texts may not be strong enough for CRFs to outperform Classification Tree. This might also be contributable to the built-in variable selection procedures of Classification Trees, which makes it more robust to interacting and interdependent features. A confirmative explanation towards the equal performances between the CRFs and the Classification Tree classifiers requires more experiments with other machine learning algorithms.

In conclusion, this paper makes the following contributions:

1. It demonstrates that an accurate tense classifier can be constructed automatically by combining off-the-shelf machine learning techniques and inexpensive linguistic features.

2. It shows that latent features (such as verb telicity, verb punctuality and temporal ordering between adjacent events) have higher utility in tense classification than the surface linguistic features.

3. It reveals that the sequential dependency between tenses of adjacent verbs in the discourse may be rather weak.

## 7 Future Work

Temporal reference is a complicated semantic domain with rich connections among the disparate features. We investigate three latent features: telicity, punctuality, and temporal ordering between adjacent verbs. We summarize several interesting questions for future research in this section. First, besides the latent features we examined in the current paper, there are other interesting latent features to be investigated under the same theme, e.g. classes of temporal expression associated with the verbs and causal relationships among disparate events. Second, currently, the latent features are obtained through manual annotation by a single annotator. In an ideal situation, multiple annotators are desired to provide the reliability of the annotations as well as reduce the noise in annotations. Thirdly, it would be interesting to examine the utility of the same latent features for classification in the opposite direction, namely, aspect marker classification for Chinese verbs in the English-to-Chinese translation scenario. Finally, following our discussion of the degree of dependencies among verb tenses in the texts, it is desirable to study rigorously the dependencies among tenses and aspect markers for verbs in extensions of the current research.

## References


James Allen. 1981. Towards a General Theory of Action and Time. Artificial Intelligence, 23(2): 123-160.

Hans Reichenbach. 1947. Elements of Symbolic Logic. Macmillan, New York, N.Y.

B. Bruce. 1972. A Model for Temporal Reference and its Application in a Question Answering System, Artificial Intelligence. Vol. 3, No. 1, 1-25.

Inderjeet Mani, James Pustejovsky and Robert Gaizauskas. 2005. The Language of Time, Oxford Press.

Wenjie Li, Kam-Fai Wong, Caogui Hong, Chunfa Yuan. 2004. Applying Machine Learning to Chinese Temporal Relation Resolution. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 582-588.

Mari Olson, David Traum, Carol Van-ess Dykema, and Amy Weinberg. 2001. Implicit Cues for Explicit Generation: Using Telicity as a Cue for Tense Structure in a Chinese to English MT System. Proceedings Machine Translation Summit VIII, Santiago de Compostela, Spain.

Yang Ye, Zhu Zhang. 2005. Tense Tagging for Verbs in Cross-Lingual Context: a Case Study. Proceedings of IJCNLP 2005, 885-895

Norbert Hornstein. 1990. As Time Goes By: Tense and Universal Grammar. The MIT Press.

James Pustejovsky, Robert Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2004. The Specification Language TimeML. The Language of Time: A Reader. Oxford, 185-96.

Zeno Vendler. 1967. Verbs and Times. Linguistics in Philosophy, 97-121.

Lafferty, J., McCallum, A. and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML-01, 282-289.

Sha, F. and Pereira, F. 2003. Shallow Parsing with Conditional Random Fields. Proceedings of the 2003 Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT/NAACL-03)

Pinto, D., McCallum, A., Lee, X. and Croft, W. B. 2003. Table Extraction Using Conditional Random Fields. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)

McCallum, A. and Li, W. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. Proceedings of the Seventh Conference on Natural Language Learning (CoNLL)

McCallum, A. K. 2002. MALLET: A Machine Learning for Language Toolkit, http://mallet.cs.umass.edu.

Jacob Cohen, 1960. A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, 20, 37-46.

Ross Ihaka and Robert Gentleman. 1996. R: A Language for Data Analysis and Graphics, Journal of Computational and Graphical Statistics, Vol. 5. 299?14.