Tense Tagging for Verbs in Cross-Lingual Context: A Case Study

Yang Ye[1] and Zhu Zhang[2]

[1] Department of Linguistics, University of Michigan, USA
[2] School of Information and Department of Electrical Engineering and Computer Science, University of Michigan, USA
yye@umich.edu, zhuzhang@umich.edu

Abstract. The current work applies Conditional Random Fields to the problem of temporal reference mapping from Chinese text to English text. The learning algorithm utilizes a moderate number of linguistic features that are easy and inexpensive to obtain. We train a tense classifier upon a small amount of manually labeled data. The evaluation results are promising according to standard measures as well as in comparison with a pilot tense annotation experiment involving human judges. Our study exhibits potential value for full-scale machine translation systems and other natural language processing tasks in a cross-lingual scenario.

1 Introduction

Temporal resolution is a crucial dimension in natural language processing. The fact that tense does not necessarily exist as a grammatical category in many languages poses a challenge on cross-lingual applications, e.g. machine translation. The fact that English tenses and Chinese aspect markers align at word level on one hand and sub-word level on the other hand poses a challenge for temporal reference distinction translation in a statistical machine translation (MT) system. A word-based alignment algorithm will not be able to capture the temporal reference distinction when mapping between Chinese and English. Being able to successfully map the temporal reference distinction in Chinese text through disparate features onto the most appropriate tenses for the parallel English text is an important criterion for good translation quality. Languages have various levels of time reference distinction representation: some have finer grained tenses than others, as typological studies have shown. When facing the unbalanced levels of temporal reference distinction between a pair of languages, we have to optimize the mapping between the two temporal systems through intelligent learning. Most machine translation systems do not have a separate temporal reference resolution module, but if we can integrate a special module into them, the temporal reference resolution of the system could be corrected accordingly and yield a better translation. Other than machine translation, in cross-lingual question answering (CLQA) with English as the target language, the ability to successfully formulate queries and maintain the temporal reference information in the original questions is desirable.

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 885 – 895, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Related Work

2.1 Temporal Reference Modeling in Cross-Lingual Scenario

The nature of being past, present or future is highly relative and hence the information contained in tenses is often referred to as temporal reference distinction. While there is a large body of research on temporal reference in formal semantics and logic as well as in other disciplines of Linguistics, works in cross-lingual temporal reference mapping remain inadequate.

Campbell et. al. [1] proposed a language-neutral framework for representing semantic tense. This framework is called the Language Neutral Syntax (LNS). Based on the observation that grammatical or morphological tenses in different languages do not necessarily mean the same thing, they interpret semantic tense to be largely a representation of event sequence; their work did not attempt direct and explicit representations of tenses. The tense node in the LNS tree contains either global tense feature (also known as "absolute tense") or anchorable tense feature (also known as "relative tense"). This work treated compound tenses as being represented by primary and secondary tense features. The tense in an embedded clause is anchored to the tense in the matrix clause. Campbell's work attempted neither a strict nor a deep semantic representation of tenses, but rather a syntactic representation that is language-neutral. In addition, similar to most of its peer works in tense modeling, it only attacked the problem in a scope of individual sentences.

Pustejovsky et. al. [2] reported an annotation scheme, the TimeML metadata for markup of events and their anchoring in documents. The challenge of human labeling of links among eventualities were discussed to the full fledge in their paper showing that inter-annotator consistency for links is a hard-to-reach ideal. The automatic "time-stamping" was attempted earlier on a small sample of text in an earlier work of Mani [3]. The result was not particularly promising showing need for bigger size of training data as well as more predictive features, especially on the discourse level. At the word level, semantic representation of tenses could be approached in various ways depending on different applications. None of the previous works were designed particularly for cross-lingual temporal reference distinction mapping and the challenges of this mapping for some language pairs have not received full attention.

2.2 Temporal Reference Mapping Between Chinese and English

Since temporal reference distinction mapping is of particular interest of cross-lingual natural language processing tasks, the pilot works for tense classification in Chinese were naturally motivated by machine translation scenario. Olsen et. al. [4] attacked tense reconstructing for Chinese text in the scenario of Chinese to English MT. On top of the more overt features, their work made 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 proposed that past tense corresponds to the telic LCS which is either inherently telic or derived telic. While grammatical aspect markings supersede the LCS, in the absence of grammatical aspect marking, verbs that have telic LCS are translated into past tense and present tense otherwise. This work, while pushing tense reconstruction one step further towards the semantics embedded in the events, is subject to the risk of adopting one-to-

one mapping between grammatical aspect markings and tenses hence oversimplifies the temporal reference situation in Chinese text. Additionally, their binary tense taxonomy is oversimplifying the rich temporal reference system that exists in Chinese.

Li et. al. [5] proposed a computational model based on machine learning and heterogeneous collaborative bootstrapping for analyzing temporal relations in a Chinese multiple-clause sentence. The core model is a set of rules that map the combinational effects of a set of linguistic features to one class of temporal relations for one event pair. Their work showed promising results for combining machine learning algorithms and linguistic features to achieve temporal relation resolution, but did not directly address cross-lingual temporal reference information mapping. The nature of the task they were attacking is B Series temporal resolution in Mctaggart's terminology.

3 Problem Definition

3.1 The Taxonomy of Tenses

In the current literature, the taxonomy of tenses typically includes the three basic tenses (present, past and future) plus their combination with the progressive and perfect grammatical aspects, because in English tense and aspect are morphologically merged. This yields a taxonomy of 13 tenses. We collapse these 13 tenses into a taxonomy of three classes: present, past and future. The reason for this collapse is twofold: linguistically, this three-class taxonomy conforms more strictly with the well defined tripartite temporal reference distinction [6]; and in practice, only nine tenses occurred in our data set: simple past, simple future, simple present, present perfect,

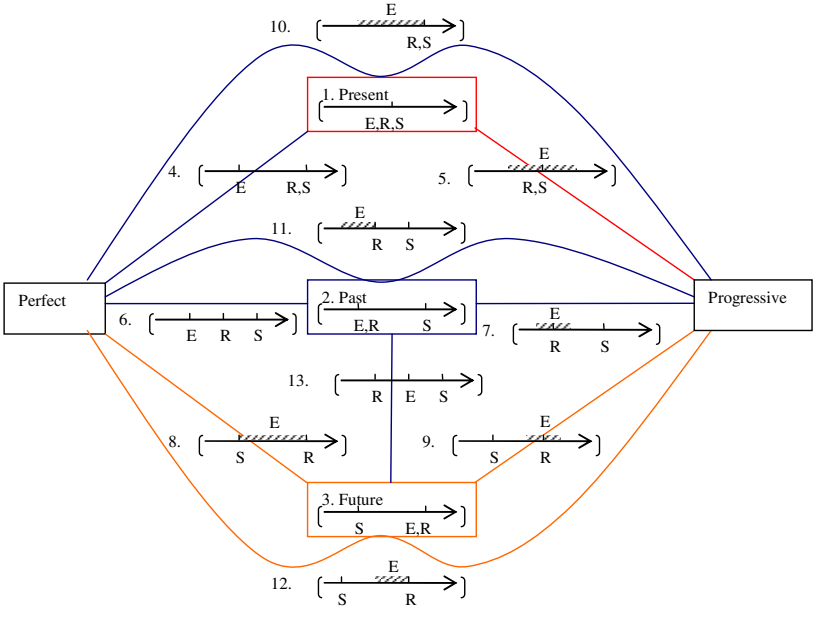

Fig. 1. Tense Taxonomy

present progressive, past perfect, past progressive, past future and present perfect progressive. Some tenses are very sparse in the data set yielding little value from the learning perspective. Figure 1[1] shows the tense taxonomy. In the graph, for each of the thirteen tenses, we provide the timeline representation for the configuration of the three time points under Reichenbachian system. E stands for the event time, R stands for the reference time and S stands for the speech time. It is observed that in terms of the relationship between the speech time and the event time, the thirteen tenses could be grouped into three categories: tense 1 and tense 5 have the event time overlapping with the speech time; tense 2, 4, 6, 7, 10, 11 and 13 have the event time being prior to the speech time; tense 3, 8, 9 and 12 have the event time being later than the speech time. These three categories form our collapsed tense taxonomy.

3.2 Problem Formulation

In general, the tense tagging problem for verbs can be formalized as a standard classification or labeling problem, in which we try to learn a classifier

$$C: V \rightarrow T$$

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

This is, however, a somewhat simplistic view of the picture. Just as temporal events are usually sequentially correlated, verbs in adjacent linguistic utterances are not independent. Therefore the problem should be further formalized as a sequential learning problem, where we try tag a sequence of verbs (V_1, …, V_n) with a sequence of tense tags (t_1, …, t_n). This formalization shares similarities with many other problems inside and outside the computational linguistics community, such as information extraction from web pages, part-of-speech tagging, protein and DNA sequence analysis, and computer intrusion detection.

4 Data

4.1 Data Summary

We use 52 pairs of parallel Chinese-English articles from LDC release. The 52 Chinese articles from Xinhua News Service consist of 20626 Chinese characters in total with each article containing between about 340 and 400 Chinese characters The Chinese documents are in Chinese Treebank format with catalog number LDC2001T11. The parallel English articles are from Multiple-Translation Chinese (MTC) Corpus from LDC with catalog number LDC2002T01. We use the best human translations out of 10 translation teams[2] as our gold-standard parallel English data.

---

[1] For tense 13, it is controversial whether the event time precedes or succeeds the speech time. (e.g. for "I was going to ask him at that time", it is not clear whether the asking event has happened by the speech time.) This graph only represents the authors' hunch about the tense taxonomy for this particular project.

[2] Two LDC personnel, one a Chinese-dominant bilingual and the other an English-dominant bilingual, performed this ranking. There was overall agreement on the ranking between the two and minor discrepancies were resolved through discussion and comparison of additional files.

4.2 Obtaining Tense Tags from the Data

The decision of the granularity level of the data points in the current project is a non-trivial issue. Recently it has been argued that tense should be regarded as a category of the whole sentence, or in logical terms of the whole proposition, since it relates to the truth value of the proposition as a whole, rather than just some property of the verb. While we agree with this assertion, in the interest of focusing on our immediate goal of assigning an appropriate tense tag to the parallel verb in the target language, we adopt the more traditional analysis of tense as a category of the verb on the basis of its morphological attachment to the verb.

There are a total of 1542 verbs in the 52 Chinese source articles. We manually aligned these verbs in the Chinese source article with their corresponding verbs in English; this yields a subset of 712 verbs out of the 1542 verbs being translated into English as verbs. We see a dramatic nominalization (i.e. verbal expressions in Chinese are translated into nominal phrases in English) process in Chinese-to-English translation through the dramatic contrast between these two numbers. We excluded the verbs that are not translated as verbs into the parallel English text. 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.

5 Tense Tagging by Learning

5.1 Temporal Reference Distinction in Chinese Text

Assigning accurate tense tags to the English verbs in Chinese-to-English Machine Translation is equivalent to understanding temporal reference distinction in the source Chinese text. Since there are no morphologically realized tenses in Chinese, the temporal reference distinction in Chinese is encoded in disparate linguistic features. Figure 2 shows how various features in simple Chinese sentences jointly represent the temporal reference distinction information. For complex sentences with an embedding structure, these features will behave in a more complicated way in that the anaphoric relations

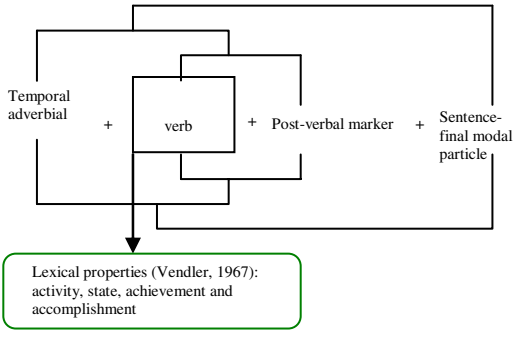

Fig. 2. Temporal Structure for a Simple Chinese Sentence

between the reference time and speech time hold differently for main verbs and verbs in embedded structure. While world knowledge is beyond the scope of our computational capacity at this stage, we expect that the various linguistic features will be able to approximately reconstruct the temporal reference distinction for Chinese verbs.

5.2 The Feature Space

There are a big variety of heterogeneous features that contribute to the temporal reference semantics of Chinese verbs. Tenses in English, while manifesting temporal reference distinction, do not always reflect the distinction at the semantic level, as is shown in the sentence "I will leave when he comes." Hornstein [7] accounted for this type of phenomenon by proposing the Constraints on Derived Tense Structures. Hence the feature space we propose to use consists of the features that contribute to the semantic level temporal reference construction as well as those contributing to the tense generation from that semantic level.

The feature space includes the following 11 features:

feature1: whether the current sentence contains a temporal noun phrase, a temporal location phrase or a temporal prepositional phrase;
feature2: whether or not the current verb is in quoted speech;
feature3: whether the current verb appears in relative clause or sentential complement;
feature4: whether or not the current verb is in news headlines;
feature5: previous word's POS;
feature6: current verb's POS, there are three types of verbs in the corpora: the regular verbs (VV); the copula "shi4"[3] (VC) and the verb "you3" (VE);
feature7: next word's POS;
feature8: whether or not the verb is followed by the aspect marker "le";
feature9: whether or not the verb is followed by the aspect marker "zhe";
feature10: whether or not the verb is followed by the aspect marker "guo";
feature11: whether or not the verb is a main verb;

The above 11 features include lexical features as well as syntactic features. None of the above features is expensive to obtain. We aim to show that the temporal reference distinction classe, as a semantic feature of the verb, could be predicted by learning from inexpensive linguistic features that are easily available. Feature 11 is motivated by the observation that tense in English is used to inform the reader (listener) of when the event associating with the main verb occurs with respect to the time of utterance while the tense of an embedded verb does not necessarily indicate this relationship directly. In the current paper, we have a different definition for main verb: any verb that is not in embedded structure is treated as a main verb including those verbs appearing in adjunct clauses.

5.3 Learning Algorithm: Conditional Random Field

Conditional Random Fields (CRF) is a formalism well-suited for learning and prediction on sequential data. It is a probabilistic framework proposed by Lafferty [8] for

[3] The digit at the end of the syllable here indicates the tone. "Shi4" means "be" and "you3" means "have".

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); 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 real-world problems, in particular, NLP problems such as shallow parsing [9], table extraction [10], and named entity recognition [11].

For our experiments, we use the off-the-shelf implementation of CRFs provided by MALLET [12].

6 Experiments and Evaluation

6.1 Preliminary Experiment with Tense Annotation by Human Judges

In order to evaluate the empirical challenge of tense generation in a Chinese-to-English Machine Translation system, a pilot experiment of tense annotation for Chinese text by native judges was carried. The annotation experiment was carried out on 20 news articles from LDC Xinhua News release with category number LDC2001T11. The articles were divided into 4 groups with 5 articles in each group. For each group, three native Chinese speakers annotated the tense of the verbs in the articles. Prior to annotating the data, the judges underwent brief training during which they were asked to read an example of a Chinese sentence for each tense and make sure they understand the examples. During the annotation, the judges were asked to read whole articles first and then select a tense tag based on the context of each verb. In cases where the judges were unable to decide the tense of a verb, they were instructed to tag it as "unknown".

Kappa scores were calculated for the three human judges' annotation results. Kappa score is the de facto standard for evaluating inter-judge agreement on tagging tasks. It is defined by the following formula (1), where P(A) is the observed agreement among the judges and P(E) is the expected agreement:

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

The annotation was originally carried out on the taxonomy of 13 tenses. We collapsed these 13 tenses into three tenses as discussed in section 3.1. Table 1 summarizes the kappa statistics for the human annotation results after we collapse the tenses:

Table 1. Kappa Scores for Human Tense Annotation for Xinhua News on Collapsed Tense Classes

	Xinhua news 1	Xinhua news 2	Xinhua news 3	Xinhua news 4
Kappa score for 3 judges	0.409	0.440	0.317	0.325

There are different interpretations as to what is a good level of agreement and what kappa scores are considered low. But generally, a kappa score of lower than 0.40 falls into the lower range of agreement[4]. Even if we consider the meta-linguistic nature of the task, the kappa scores we observe belong to the poor-fair range of agreement, illustrating the challenge of temporal reference mapping across Chinese and English. The difficulty of tense classification demonstrated by these experiments with human judges provides an upper bound on the performance of automatic machine classification. As challenging a task as it is, tense generation for English verbs in a Chinese-to-English Machine Translation system must address this cross-lingual mapping problem in order to obtain an accurate translation result.

6.2 Experimental Setup and Evaluation Metrics

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. Based on this conjecture, we experiment with two different sequential granularities:

— Sentence-level sequence: each sentence is treated as a sequence;
— Paragraph-level sequence: each sentence is treated as a sequence, and there is no boundary between sentences within the paragraph.

All results are obtained by 5-fold cross validation. The classifier's performance is evaluated against the tenses from the best-ranked human translation parallel English text.

To evaluate the performance of classifiers, we measure the standard classification accuracy where accuracy is defined as in equation (2):

$$accuracy = \frac{number\quad of\quad correct\quad predictions}{total\quad number\quad of\quad predictions} \tag{2}$$

To measure how well the classifier does on each class respectively, we compute precision, recall, and F-measure, which are defined respectively in equation (3), (4) and (5):

$$\Pr ecision = \frac{number\quad of\quad correct\quad hits}{total\quad number\quad of\quad hits} \tag{3}$$

$$\mathrm{Re}\, call = \frac{number\quad of\quad correct\quad hits}{size\quad of\quad perfect\quad hitlist} \tag{4}$$

$$F - measure = \frac{2 \times \Pr ecision \times \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \tag{5}$$

[4] http://www.childrens-mercy.org/stats/definitions/kappa.htm

6.3 Experimental Results

The evaluation is carried on the collapsed tense taxonomy that consists of three tense classes: present, past and future. This collapse is motivated by two reasons: linguistically, this collapse reflects the accommodation of the "gray area" that exists in the 13-way tense taxonomy; practically, the collapse helps to alleviate the sparse data problem. Ideally, with a large enough data set that could cover the less-common tenses, the full-fledged tense taxonomy is desirable given that the "gray area" could be analyzed and included into the evaluation. The CRF-based tense classifier yielded the performance in Table 2 and Table 3:

Table 2. Sentence-level sequence: overall accuracy 58.21%

	Precision	Recall	F-measure
Present	42.50%	27.48%	32.07%
Past	67.57%	79.55%	72.10%
Future	29.66%	25.56%	21.56%

Table 3. Paragraph-level sequence: overall accuracy 58.05%

	Precision	Recall	F-measure
Present	38.79%	32.44%	33.96%
Past	69.12%	75.72%	71.59%
Future	33.16%	30.25%	26.59%

An accuracy of around 60% seems not satisfactory if viewed in isolation, but when contrasted with the kappa score of human tense annotating discussed above, the current evaluation indicates promising results for our algorithm. Even though the human judges underwent only minimal training, their poor-to-fair kappa scores indicate that this is a very hard problem. Therefore while there is certainly room for improvement, the tagging performance of our algorithm is quite promising.

It is noticed that the granularity of sequences does not seem to yield significantly different performance based on the current data. However, whether this is true in general remains an open question.

7 Discussions

There are four important dimensions for any natural language processing tasks:

— The data: ideally, the data used should be as representative as possible of a wide range of genres unless the target application is focused on a certain narrow domain;
— The feature space: ideally, the features should be easily available and have wide coverage over the predicting space of the target problem; the more so-

phisticated and the more expensive the features are, the less we could claim to gain from the learning algorithms.
— The learning algorithm: nowadays, various machine-learning algorithms have been proposed and applied in different natural language task domains. A learning algorithm should be chosen to appropriately explore the feature space.
— The evaluation: ideally, evaluation from multiple perspectives is desired to resolve disagreements.

Reflecting upon these dimensions for the current paper, from the data perspective, we focused on news report genre where the temporal thread progression is relatively simpler than many other genres. When facing temporal reference classification for more complicated genres, larger amounts of training data would be necessary for learning a more sophisticated classifier. Fortunately, the amount of accessible parallel data is growing and it is always possible to obtain the tense tags for the Chinese verbs automatically using an off-the-shelf aligning tool although this might introduce a certain amount of noise.

As for the choice of the predicting features, the current project does not utilize any lexical semantic features owing to the limited lexical semantic knowledge resources for Chinese. We expect such knowledge resources, if available, would enhance the feature vector and boost the classification performance. Additionally, it is observed that for a Chinese-to-English MT system, tense generation in English is significantly subject to the syntactic constraints. Hence when integrating into a MT system, the current learning algorithm might have opportunity to employ additional features from other parts of the system, for example, syntactic features for English could be added to the current feature space.

Regarding the choice of learning algorithm, we chose CRFs, a learning algorithm for sequential data, based on the fact that tenses for verbs in a certain discourse unit are not independent of each other.

From the evaluation point of view, the current work evaluates the classifier against the tenses from a certain human translation team. The frequent disagreements among the human annotators illustrate the difficulty of constructing a gold standard against which to evaluate the performance of our classifier. Lastly, measuring BLEU score change brought about by integrating the current classifier into a statistical MT system would be desirable, such that we can better understand the practical implications of this study for MT systems.

8 Conclusions and Future Work

The current work has shown how a moderate set of shallow and inexpensive linguistic features can be combined with a standard machine learning algorithm for learning a tense classifier trained on a moderate number of data points, with promising results. A tense resolution module built upon the current framework could enhance a MT system with its temporal reference distinction resolution.

Several issues to be explored in future work are the following: First, our current training corpus of Xinhua News articles is rather homogeneous, hence the classifier trained exclusively on this data set may not be robust when carried over to data from different source. This will become particularly important if we want to integrate the

current work into a general-domain MT system. Secondly, related to the homogeneity of our training data, we only explored a limited number of features, while the feature space could be expanded to include a richer and wider scope. For example, discourse structure features have not been explored. Finally, we are very interested in evaluating our work against existing MT systems with regard to temporal mapping.

References

1. Campbell, R., Aikawa, T., Jiang, Z., Lozano, C., Melero, M and Wu, A.: A Language-Neutral Representation of Temporal Information. In Proceedings of the Workshop on Annotation Standards for Tempora Information in Natural Language, LREC 2002, Las Palmas de Gran Canaria, Spain (2002) 13-21.
2. Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G. and Mani, I.: The Specification Language TimeML. In Mani, I., Pustejovsky, J., and Gaizauskas, R (eds.). (2004) The Language of Time: A Reader. Oxford University Press, to appear
3. Mani, I.: "Recent Developments in Temporal Information Extraction (Draft)", In Nicolov, N., and Mitkov, R. Proceedings of RANLP'03, John Benjamins, to appear.
4. Olson, M., Traum, D., Van-ess Dykema, C. and Weinberg, A.: Implicit Cues for Explicit Generation: Using Telicity as a Cue for Tense Structure in a Chinese to English MT System, in proceedings Machine Translation Summit VIII, Santiago de Compostela (Spain) (2001)
5. Li, W., Wong, K. F., Hong, C. and Yuan, C.: Applying Machine Learning to Chinese Temporal Relation Resolution, Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (2004) 582-588
6. Reichenbach, H.: Elements of Symbolic Logic, The Macmillan Company (1947)
7. Dorr, B. J. and Gaasterland, T.: "Constraints on the Generation of Tense, Aspect, and Connecting Words from Temporal Expressions," Technical Report CS-TR-4391, UMIACS-TR-2002-71, LAMP-TR-091, University of Maryland, College Park, MD (2002)
8. Lafferty, J., McCallum, A. and Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML-01, (2001) 282-289
9. Sha, F. and Pereira, F.: Shallow Parsing with Conditional Random Fields, In Proceedings of the 2003 Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT/NAACL-03) (2003)
10. Pinto, D., McCallum, A., Lee, X. and Croft, W. B.: Table Extraction Using Conditional Random Fields. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003) (2003)
11. McCallum, A. and Li, W.: Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In Proceedings of the Seventh Conference on Natural Language Learning (CoNLL) (2003)
12. McCallum, A. K.: MALLET: A Machine Learning for Language Toolkit http://mallet.cs.umass.edu. (2002)