

Classifier-based Tense Model for SMT

GONG ZhengXian¹ ZHANG Min² TAN ChewLim³ ZHOU GuoDong^{1*}

(1) School of Computer Science and Technology, Soochow University, Suzhou, China 215006

(2) Human Language Technology, Institute for Infocomm Research, Singapore 138632

(3) School of Computing, National University of Singapore, Singapore 117417

{zhxgong, gdzhou}@suda.edu.cn, mzhang@i2r.a-star.edu.sg,
tancl@comp.nus.edu.sg

ABSTRACT

Tense of one sentence can indicate the time when an event takes place. Therefore, it is very useful for natural language processing tasks such as Machine Translation (MT). However, the mapping of tense in MT is a very challenging problem as the usage of tenses varies from one language to another. Aiming at translating one language (source) which lacks overt tense markers into another language (target) whose tense information is easily recognized, we propose to use a classifier-based tense model to keep the main tense in target side consistent with the one in source side. Furthermore, we present a simple and effective way to help this model by expanding more phrase pairs with different tenses. Experimental results demonstrate our methods significantly improve translation accuracy.

TITLE AND ABSTRACT IN ANOTHER LANGUAGE (CHINESE)

基于分类的时态模型在SMT里的应用

时态对一些自然语言处理任务来说特别重要, 比如机器翻译. 然而, 在翻译过程中要保持源、目标端的时态一致性是非常困难的, 尤其表现在对某些没有明显时态标记的源语言句子进行翻译的时候. 本文提出了一种基于分类技术的时态模型去帮助这类翻译. 此外, 本文还采用了一种简单有效的方法来获取更多时态的短语对. 实验表明我们的方法能够显著提高翻译质量.

KEYWORDS: Statistical Machine Translation(SMT), Tense Model, Verb Phrase .

KEYWORDS IN CHINESE: 统计机器翻译, 时态模型, 动词短语 .

*Corresponding author.

1 Introduction

Correct usage of tenses is important because they encode the temporal order of events in a text and odd tense can lead to confusion and misunderstanding in communication. However, there is a big difference in the usage of tenses in different languages. In inflectional languages like English, tense is often expressed by verb inflections and thus can be easily recognized. Whereas, some of the major Eastern Asian languages such as Chinese, Vietnamese and Thai, do not have the grammatical category of tense, and their tense is indicated by content words such as adverbs of time. So mapping a correct tense from source-side into target-side in this case is difficult and thus it poses challenges on current machine translation tasks.

In some Interlingua-based MT systems (Dorr, 1992; Olsen et al., 2001; Wang and Seneff, 2006), tense information of the source language can be firstly transformed into an abstract language-independent representation and passed to the target language. However, such research works have not been thoroughly studied since the definition of an Interlingua is already very difficult esp. for a wider domain. With the popularity of SMT, corpus-based methods of addressing tense problems for MT have been introduced (Ye and Zhang, 2005; Liu et al., 2011; Lee, 2011). However, these works just resolve tense recognition and do not address how to integrate their works into a realistic SMT system.

There is little work on resolving tense error for SMT. Gong et al. (2012) propose target N-gram-based tense models to improve translation performance. However, it is not reliable enough since they only consider the target-side tense information and SMT systems often generate abnormal outputs. For the source language with weak inflections or even without inflections, its contexts can provide valuable cues about tense. For example, some Chinese words, such as “经常(JinChang, often)” and “昨天(ZuoTian, yesterday)”, can obviously indicate present tense and past tense respectively.

Our proposed SMT system, both source-side and target-side tense cues are employed. Aiming at translating one language (source) which lacks overt tense markers into another language (target) whose tense information is easily recognized, we first automatically construct two tense classifiers based on a Chinese-to-English parallel corpus. Then, two related features are specially designed for a phrase-based SMT system. Furthermore, since verb phrases have limited tense forms in our phrase table, we propose a simple and effective solution to expand phrase table. Experimental results on Chinese-to-English translation show that classifier-based tense model can obtain significant improvements over the baseline.

2 SMT with classifier-based tense model

2.1 Basic idea

In Chinese-to-English translation, the taxonomy of English tenses typically includes three basic tenses (present, past and future) plus their combination with the progressive and perfect grammatical aspects. Since Chinese sentences lack tense markers, it is natural to label them with correct tenses before translation. We treat this labeling task as a classification problem and train a multi-class SVM classifier to assign four labels: P_r - present tense; P_a -past tense; F -future tense; UNK -unknown tense. Given T_s is the major tense of one Chinese sentence f ,

$$P(T_s|f) = \arg \max P(T_{s_i}|f), \quad i = 1 \text{ to } 4, \quad T_{s_i} \in \{P_r, P_a, F, UNK\}.$$

where $P(T_{s_i}|f)$ is the probability that the major tense of f equals to T_{s_i} . Furthermore, given e_{best} is the most possible equivalent translation of f and T_g is the major tense of e_{best} , we confine

our translation model to satisfy this condition: $T_g = T_s$. That means we assume hypothesis translations with good quality should keep the same tense with source-side sentence.

This idea is straightforward but its implementation is difficult. First of all, source sentences lack inflections and it brings difficulty in determining T_g . Furthermore, there exists the same degree of difficulty in determining T_g for SMT outputs because they are uncompleted texts with abnormal word ordering (Vilar et al., 2006). We treat such task of determining T_g and T_s as tense predication. Although our source-side tense predication is slightly similar to Liu et al. (2011), our predication is special in two aspects: (1) we only consider the major tense of source-side sentence and thus our classification accuracy is high enough; (2) we exploit more useful features for this task. Furthermore, we integrate our tense model into a popular SMT system and improve the translation results.

2.2 The system framework

The work described in this paper is based on a modified Moses, a state-of-the-art phrase-based SMT system. The major modified parts for Moses are input and output modules in order to translate using document-level information. Our SMT system follows Koehn et al. (2003) and adopts similar six groups of features. Besides, the log-linear model (Och and Ney, 2000) is employed to linearly interpolate these features according to formula(1):

$$e_{best} = \arg \max_e \sum_{m=1}^M \lambda_m h_m(e, f) \quad (1)$$

where $h_m(e, f)$ is a feature function, and λ_m is the weight of $h_m(e, f)$ optimized by a discriminative training method on a held-out development data (Och, 2003).

Classifier-based tense model can be easily integrated into the formula(1) by the following special features:

$$F_1 = \begin{cases} 1 & \text{if } (t_g = t_s) \\ 0 & \text{else} \end{cases} \quad F_2 = P(T_{s_i} | f)$$

F_1 is a binary feature which encourages the decoder to prefer hypothesis translations which have tense form conforming to the source-side contexts. Since tense of source-side sentence sometimes is not reliable enough, F_2 is introduced to inform the decoder to what extent it can trust F_1 , it indicates the confidence of the source-side tense classifier. It is worth noting that we don't introduce the similar feature for the target side since the classification accuracy for hypothesis translation is low (see section 3.3).

Before translation, our decoder uses a trained source-side tense classifier to predict its major tense and obtains the value of F_2 . During decoding, when a hypothesis translation covers the whole source-side words, the decoder will use a trained target classifier to predict its major tense and compute F_1 . The two feature values with related tuned weights are used to re-rank all hypothesis translations.

3 Constructing tense classifiers

3.1 Prepare training data

The key of constructing tense classifiers is to obtain a collection of labelled data. Because of lack of annotated training data, we propose to generate reference tenses based on a Chinese-English parallel corpus (FBIS, see Section 5).

To English sentences in this parallel corpus, we first automatically identify their major tenses according to the morphology of their *root word*. Here *root word* is the word with typed dependency of “root” in a dependency parsing tree obtained from the Stanford dependencies¹. If *root word* is a verb with special POS tag^{2,3}, the major tense can be easily recognized in most cases. If *root word* can not determine the specific tense, we will traverse the dependency tree forward and backward to find other verb governed by *root word* directly. The worst case is to use “UNK” as the major tense for an English sentence.

Since we use parallel corpus, it is reasonable that we treat such major tenses produced by English sentences as gold standards to train a classifier for their corresponding Chinese sentences.

3.2 Source-side tense classifier

After obtaining training data, we built a 4-class (p_r, p_a, E, UNK) tense classifier for source-side (Chinese) sentences with the tool of SVM^{multiclass}⁴ accompanied by the following features:

(1) *Words and POS patterns (WP) features*: Combinations of word/POS tag for each of words in the whole sentence. These features are expected to capture some special expression patterns, for example, a Chinese verb followed by the word “了(Le)” often refers to past tense.

(2) *Temporal word feature*: The word with the typed dependency of “tmod” (means temporal modifier). A complicated sentence maybe contains multiple “tmod” words. We only consider such “tmod” word who is governed by its “major” verb directly. Here the major verb maybe a *root word* or a verb governed by *root word* directly. This feature can catch special temporal words, such as “近日(JinRi, recently)”, “今年(JinNian, this year)”.

(3) *History tense feature*: History tense means tense of previous sentence. We found some Chinese sentences have no temporal words or markers at all and thus their limited contexts at sentence level can not classify them with correct tense form by itself. Gong et al. (2012) and Lee (2011) show history tense has close relation to current sentence.

(4) *The category of document feature*: This feature is expected to control the feature of “history tense”. For example, articles about products and laws like to use present tense and in this case history tense is important. However, the importance of history tense in other domains such as news reports may be reduced since they tend to use tense diversely. In order to obtain the category of documents, another SVM classifier is trained based on a public classification corpus⁵, which only uses bag of words and TF/IDF (Salton et al., 1975) as feature/value.

The first two features refer to lexical information at sentence level and the latter two features reflect high-level semantic at document level.

After constructing a source-side classifier, we use classification accuracy to measure its performance. Table 1 shows the 5-fold cross validation results using different features for Chinese sentences from previous training set. The performance is improved incrementally by adding above features. When “WP” features are used, the accuracy is 75.44%. Adding “Temporal

¹<http://nlp.stanford.edu/software/lex-parser.shtml>

²POS tags can distinguish five different forms of verbs: the base form (tagged as VB), and forms with overt endings D for past tense, G for present participle, N for past participle, and Z for third person singular present. It is worthy to note that VB, VBG and VBN cannot determine the specific tenses by themselves.

³special modal verbs with POS tag “MD”, such as “will”, “shall”, “ll”, indicate future tense

⁴http://svmlight.joachims.org/svm_multiclass.html

⁵http://www.nlp.org.cn/docs/doclist.php?cat_id=16&type=15

word” feature obtains a gain of 3.06%. After introducing “history tense” feature, a slightly improvement with 1.59% is yielded. Finally, the classification accuracy reaches to 83.10% by adding “category of doc” feature. It is obvious that the “history tense” feature is largely affected by different document categories.

Table 1 also shows one result for Chinese sentences from a test set “NIST 2005”(see Section 5). There are 4 English references for NIST 2005, we choose one of them and determine their major tense using the method described in Section 3.1 and thus obtain the gold standards for “NIST 2005”. we use the source-side tense classifier to predict tenses for NIST 2005, the accuracy of this 4-class classifier can reach to 78.26.

Data	Features	Accuracy(%)
Training Data	WP	75.44
	+Temporal word	78.50
	+History tense	80.09
	+Category of DOC	83.10
NIST 2005	ALL	78.26

Table 1: Classification accuracy for Chinese sentence with different features

Data	Features	Accuracy(%)
Training Data	Words	81.46
	+POS	86.22
Reference of NIST 2005	Words	79.01
	+POS	81.00
1-best translation of NIST 2005	Words	59.07
	+POS	57.02
Oracle translation of NIST 2005	Words	66.40
	+POS	64.70

Table 2: Classification accuracy for English sentence with different features.

3.3 Target tense classifier for SMT outputs

Unlike normal English texts in previous training data, the SMT outputs are very noisy. So we can not determine their major tense using the method described in Section 3.1. Similarly we try to build a tense classifier for these “special” English sentences based on the previous training set only according to two kinds of feature, words and POS tags.

We first measure the cross validation performance for this target-side tense classifier. We can obtain the classification accuracy of 81.46% (Shown on Table 2, 5-fold cross validation) only with “words” features. Then, we use this classifier to classify one reference in “NIST 2005” and the accuracy can still keep to 79.01%. However, when we measure on the translation results (1-best) produced by our baseline, the classification accuracy sharply drops to 59.07%. It is interesting that the classifier’s accuracy on oracle translation texts (hypothesis translations with the highest BLEU score from the N-best lists) can rise to 66.40%. From these experimental results, we can conclude that the loss of target-side tense classification accuracy is due to the imperfect translations.

It is worth noting that POS tags can contribute to predict tense for normal English sentence but slightly harm the performance for SMT outputs since POS Tagger can give wrong POS tags by automatically adjusting to its current contexts (Och et al., 2004). In order to use more correct POS tags, we pass the original POS tags to our decoder by introducing an expanded phrase table described in Section 4.1.

4 Expanding verb phrases

Another important impact factor of employing tense model in a phrase-based SMT system is phrase pairs. The core of a phrase-based statistical machine translation system is a phrase table containing pairs of source and target language phrases, each weighted by a conditional translation probability. Koehn et al. (2003) showed that translation quality is very sensitive

to this table. If the required phrase pairs with correct tense forms are not in phrase table, the tense model will never play an important role even it is effective in principle. The ideal way to address this problem is to build large parallel corpus but the cost is huge. In this section, we propose a simple and effective way to expand some special phrase pairs.

4.1 Expanding procedure

The detailed procedure is described as follows,

1. POS tagging the English sentences in parallel corpus. Then performing word alignment and phrase extraction between normal Chinese sentences and special Word/POS sentences. So one phrase table, denoted by **PT**, whose target phrases containing POS tags is generated;
2. Traversing each phrase pair whose target phrases contains verb (verb phrases in short). We denote such phrase pair by ph , the source phrase of ph by ph_{sc} , and the target phrase of ph by ph_{tg} ;
3. Searching verb word of ph_{tg} , and automatically generating new verbs with other tense forms according to serials of transformation rules. For each new verb, expanding a new target phrase by associating other words of ph_{tg} . Putting the new target phrase together with ph_{sc} , POS tags and some translation probabilities of ph into a new table **PA**.
4. Generating an expanded phrase table **PE** by merging **PA** and **PT**.

During translation, our decoder only employs the phrase pair and POS tags of **PE** and discard the translation probability parts of **PE** for reducing the influence of data sparsity.

4.2 Automatic transformation rules for verb phrases

The transformation rules described in Section 4.1 involve the following directions: past \rightarrow present, present \rightarrow past, base form \rightarrow past and base form \rightarrow future. For example, “was *” is evolved into “is *”; “referred to *” into “refer(s) to *”; “operate *” into “will operate *”. When doing the transformation of past \rightarrow present, we need to specially consider the form of the third person singular. The transformation of present \rightarrow past involves regular verbs and irregular verbs. To irregular past, we manually construct a special mapping file (put \rightarrow put).

Some special source phrases containing words like “正(Zheng)”, “过(Guo)”, “着(Zhe)”, “了(Le)” should not be expanded with a new tense form since they have already indicated certain tenses. Furthermore, if there are two or more verbs in one phrase and all of them have the same tense form, we assure to expand all of them with one kind of new tense form.

It is worth noting some useful verb phrases are produced by expanding rules, but at the same it also brings some noisy data. For example, we wrongly derive a new target phrase “yesterday night, they discuss” for the phrase pair of “昨晚 商讨(ZuoWan ShangTao) ||| yesterday night, they discussed”. Maybe we can rule out them by using previous target-side tense classifier but it will cost time since phrase tables are often huge.

5 Experimentation

5.1 Experimental setting for SMT

In our experiment, SRI language modeling toolkit (Stolcke et al., 2002) was used to train a 5-Gram general language model on the Xinhua portion of the Gigaword corpus. Word alignment

Corpus		Sentences	Documents
Role	Name		
Train	FBIS	228455	10000
Dev	NIST2003	919	100
Test	NIST2005	1082	100

Table 3: Corpus statistics

was performed on the training parallel corpus using GIZA++ (Och and Ney, 2000) in two directions. We use FBIS as the training data, the 2003 NIST MT evaluation test data as the development data, and the 2005 NIST MT test data as the test data. Table 3 shows the statistics of these data sets (with document boundaries annotated).

5.2 Translation results

For evaluation, the NIST BLEU script (version 13) with the default setting is used to calculate the BLEU score (Papineni et al., 2002), which measures case-insensitive matching of 4-grams. We conduct significance tests using the paired bootstrap method (Koehn, 2004)⁶. In this paper, “***” means significantly better with p-value of 0.05. In addition, we also evaluate translation quality with METEOR (Banerjee and Lavie, 2005) and three edit-distance style metrics, Word Error Rate(WER), Position independent word Error Rate(PER) (Tillmann et al., 1997), and Translation Edit Rate(TER) (Snover et al., 2006).

The main goal of this experiment is to testify the influence of integrating the proposed tense model. From the results shown on Table 4, the systems with tense model significantly outperform the ones without it. The system with tense model (Baseline+Tense Model) yields a gain of 0.74 in BLEU score compared to the baseline and the METEOR score rises to 53.11. In all cases, all edit-distance errors are reduced.

The other goal is to see whether expanded phrase table can contribute to SMT system or not. After being expanded, the size of phrase table has a rise of 35%. But such a big rise only bring a slight improvement with 0.16 in BLEU score in the system(Baseline+Ex_phrases). As our expectation, the new phrase table can help our proposed system (Baseline+Tense Model+Ex_phrases) to gain a rise about 1 point (0.97) in BLEU score. It seems our proposed system can give more chance to expanded phrase pairs.

System	BLEU	METEOR	WER	PER	TER
Moses_Md(Baseline)	28.30	52.07	65.66	43.25	59.76
Baseline+Ex_phrases	28.46	52.13	65.48	42.41	59.30
Baseline+Tense Model	29.04(***)	53.11	63.09	40.12	55.82
Baseline+Tense Model + Ex_phrases	29.27(***)	54.50	60.82	36.03	52.37

Table 4: Translation Results

5.3 Tense accuracy of SMT outputs

This experiment is designed to analyze quality of SMT outputs from a new viewpoint. We enforce the decoder to output the POS tag. Then we predict the major tense of the new SMT

⁶<http://www.ark.cs.cmu.edu/MT>

outputs with previous target-side tense classifier (see Section 3.3). We found the classification accuracy shown on table 5 is improved in both 1-best translation and oracle translation. It also shows the quality of translations has been improved due to our proposed tense model.

System	Data	Accuracy(%)
Baseline	1-best translation of NIST 2005	57.02
	Oracle translation of NIST 2005	64.70
Our best system	1-best translation of NIST 2005	68.59
	Oracle translation of NIST 2005	71.13

Table 5: Tense classification results for the best proposed system

5.4 Discussion

Our tense model is a discriminative model with rich features. The two tense features in our SMT system seem straightforward, but when the decoder uses these classifiers to keep target-side major tense consistent with source-side tense, all lexicon and shallow syntactic information in target side have been employed.

Table 6 shows an example produced by the baseline and our proposed system respectively. The source-side tense classifier predicts that the Chinese sentence has 84.39% probability to utilize past tense. The target-side tense classifier predicates the output of baseline and our system is in past tense with the probability of 53.01% and 78.11% respectively. We think both occurrence of “was” and “is” in the baseline confuses the classifier.

Src	这个组织当年是在赫尔辛基举行的会议成立,对结束冷战有其贡献。
Ref	The organization was established at the Helsinki conference,which helped end the Cold War.
Baseline	the organization was held for years helsinki conference,is a contribution to end the cold war.
Ours	the organization in those days was held at the helsinki conference,a contribution to end the cold war.

Table 6: An example produced by the baseline and our proposed system respectively

Conclusion and perspectives

We have incorporated source-side and target-side tense knowledge into a phrase-based SMT system with two special features. We explore and capture more useful features to predict source-side tense. And the translation model is enhanced with an expanded phrase table which has more verb phrases with diverse tense forms. We evaluate translation results with multiple popular metrics, and especially with tense classification accuracy which can be introduced to current automatic evaluation metrics. All experimental results show our proposed system can obtain significant improvements over the baseline.

In the future, we will extend our work to consider aspect information since it can indicate the situation of events and more semantics.

Acknowledgments

This research is supported partly by NUS FRC Grant R252-000-452-112,the National Natural Science Foundation of China under grant No.61273320 and 61003155, the National High Technology Research and Development Program of China (863 Program) under grant No.2012AA011102, and the Preliminary Research Project of Soochow University.

References

- Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.
- Dorr, B. (1992). A parameterized approach to integrating aspect with lexical-semantics for machine translation. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, pages 257–264. Association for Computational Linguistics.
- Gong, Z., Zhang, M., Tan, C., and Zhou, G. (2012). N-gram-based tense models for statistical machine translation. In *Proceedings of EMNLP*, pages 276–919.
- Koehn, P., Och, F., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.
- Lee, J. (2011). Verb tense generation. *Procedia-Social and Behavioral Sciences*, 27:122–130.
- Liu, F., Liu, F., and Liu, Y. (2011). Learning from chinese-english parallel data for chinese tense prediction. In *Proceedings of IJCNLP*, pages 1116–1124, Chiang Mai, Thailand.
- Och, F. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.
- Och, F., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., et al. (2004). A smorgasbord of features for statistical machine translation. In *Proceedings of HLT-NAACL*, pages 161–168.
- Och, F. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics.
- Olsen, M., Traum, D., Van Ess-Dykema, C., Weinberg, A., et al. (2001). Implicit cues for explicit generation: using telicity as a cue for tense structure in a chinese to english mt system.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.
- Snover, M., Dor, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231, Boston.
- Stolcke, A. et al. (2002). Srilm-an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*, volume 2, pages 901–904.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated dp based search for statistical translation. In *European Conf. on Speech Communication and Technology*, pages 2667–2670, Rhodes, Greece.

Vilar, D., Xu, J., d'Haro, L., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of LREC*, pages 697–702.

Wang, C. and Seneff, S. (2006). High-quality speech-to-speech translation for computer-aided language learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 3(2):1–21.

Ye, Y. and Zhang, Z. (2005). Tense tagging for verbs in cross-lingual context: A case study. *Natural Language Processing–IJCNLP 2005*, pages 885–895.