

# N-gram-based Tense Models for Statistical Machine Translation

Zhengxian Gong<sup>1</sup> Min Zhang<sup>2</sup> Chewlim Tan<sup>3</sup> Guodong Zhou<sup>1\*</sup>

<sup>1</sup> School of Computer Science and Technology, Soochow University, Suzhou, China 215006

<sup>2</sup> Human Language Technology, Institute for Infocomm Research, Singapore 138632

<sup>3</sup> 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 is a small element to a sentence, however, error tense can raise odd grammars and result in misunderstanding. Recently, tense has drawn attention in many natural language processing applications. However, most of current Statistical Machine Translation (SMT) systems mainly depend on translation model and language model. They never consider and make full use of tense information. In this paper, we propose n-gram-based tense models for SMT and successfully integrate them into a state-of-the-art phrase-based SMT system via two additional features. Experimental results on the NIST Chinese-English translation task show that our proposed tense models are very effective, contributing performance improvement by 0.62 BLUE points over a strong baseline.

## 1 Introduction

For many NLP applications, such as event extraction and summarization, tense has been regarded as a key factor in providing temporal order. However, tense information has been largely overlooked by current SMT research. Consider the following example:

**SRC:** 这项禁运是冷战的结果,不能反映现在的情势,更未反映中国与欧盟间的伙伴关系。

**REF:** The embargo **is** a result of the Cold War and **does** not reflect the present situation nor the partnership between China and the EU.

**MOSES:** the embargo **is** the result of the cold war, not reflect the present situation, it **did** not reflect the partnership with the european union.

\*Corresponding author.

Although the translated text produced by Moses<sup>1</sup> is understandable, it has very odd tense combination from the grammatical aspect, i.e. with tense inconsistency (*is/does* in REF vs. *is/did* in Moses). Obviously, slight modification, such as changing “is” into “was”, can much improve the readability of the translated text. It is also interesting to note that such modification can much affect the evaluation. If we change “did” to “does”, the BLEU-4 score increases from 22.65 to 27.86 (as matching the 4-gram “does not reflect the” in REF). However, if we change “is” to “was”, the BLEU score decreases from 22.65 to 21.44.

The above example seems special. To testify its impact on SMT in wider range, we design a special experiment based on the 2005 NIST MT data (see section 6.1). This data has 4 references. We choose one reference and modify its sentences with error tense<sup>2</sup>. After that, we use other 3 references to measure this reference. The modified reference leads to a sharp drop in BLEU-4 score, from 52.46 to 50.27 in all. So it is not a random phenomenon that tense can affect translation results.

The key is how to detect tense errors and choose correct tenses during the translation procedure. By carefully comparing the references with Moses output, we obtain the following useful observations,

**Observation(1):** to most simple sentences, coordinate verbs should be translated with the same tense while they have different tense in Moses output;

**Observation(2):** to some compound sentences,

<sup>1</sup><http://www.statmt.org/moses/>

<sup>2</sup>Such changes are small by mainly modifying one auxiliary verb for a sentence, such as “*is* → *was*”, “*has* → *had*”.

the subordinate clause should have the consistent tense with its main clause while Moses fails;

**Observation(3)**: the diversity of tense usage in a document is common. However, in most cases, the neighbored sentences tends to share the same main tense. In some extreme examples, one tense (past or present), even dominates the whole document.

One possible solution to model above observations is using rules. Dorr (2002) refers to six basic English tense structures and defines the possible paired combinations of “present, past, future”. But the practical cases are very complicated, especially in news domain. There are a lot of complicated sentences in news articles. Our preliminary investigation shows that such six paired combinations can only cover limited real cases in Chinese-English SMT.

This paper proposes a simple yet effective method to model above observations. For each target sentence in the training corpus, we first parse it and extract its tense sequence. Then, a target-side tense n-gram model is constructed. Such model can be used to estimate the rationality of tense combination in a sentence and thus supervise SMT to reduce tense inconsistency errors against **Observations (1)** and **(2)** in the sentence-level. In comparison, **Observation (3)** actually reflects the tense distributions among one document. After extracting each main tense for each sentence, we build another tense n-gram model in the document-level. For clarity, this paper denotes document-level tense as “inter-tense” and sentence-level tense as “intra-tense”.

After that, we propose to integrate such tense models into SMT systems in a dynamic way. It is well known there are many errors in the current MT output (David et al., 2006). Unlike previously making trouble with reference texts, the BLEU-4 score cannot be influenced obviously by modifying a small part of abnormal sentences in a static way. In our system, both inter-tense and intra-tense model are integrated into a SMT system via additional features and thus can supervise the decoding procedure. During decoding, once some words with correct tense can be determined, with the help of language model and other related features, the small component—“tense”—can affect surrounding words and improve the performance of the whole sentence. Our experimental results (see the examples in Sec-

tion 6.4) show the effectiveness of this way.

Rather than the rule-based model, our models are fully statistical-based. So they can be easily scaled up and integrated into either phrase-based or syntax-based SMT systems. In this paper, we employ a strong phrase-based SMT baseline system, as proposed in Gong et al. (2011), which uses document as translation unit, for better incorporating document-level information.

The rest of this paper is organized as follows: Section 2 reviews the related work. Section 3 and 4 are related to tense models. Section 3 describes the pre-processing work for building tense models. Section 4 presents how to build target-side tense models and discuss their characteristics. Section 5 shows our way of integrating such tense models into a SMT system. Section 6 gives the experimental results. Finally, we conclude this paper in Section 7.

## 2 Related Work

In this section, we focus on related work on integrating the tense information into SMT. Since both inter- and intra-tense models need to analyze and extract tense information, we also give a brief overview on tense prediction (or tagging).

### 2.1 Tense Prediction

The tense prediction task often needs to build a model based on a large corpus annotated with temporal relations and thus its focus is on how to recognize, interpret and normalize time expressions. As a representative, Lapata and Lascarides (2006) proposed a simple yet effective data-intensive approach. In particular, they trained models on main and subordinate clauses connected with some special temporal marker words, such as “after” and “before”, and employed them in temporal inference.

Another typical task is cross-lingual tense prediction. Some languages, such as English, are inflectional, whose verbs can express tense via certain stems or suffix, while others, such as Chinese often lack inflectional forms. Take Chinese to English translation as example, if Chinese text contains particle word “了(Le)”, the nearest Chinese verb prefers to be translated into English verb with the past tense. Ye and Zhang (2005), Ye et al. (2007) and Liu et al. (2011) focus on labeling the tenses for keywords in

source-side language.

Ye and Zhang (2005) first built a small amount of manually-labeled data, which provide the tense mapping from Chinese text to English text. Then, they trained a CRF-based tense classifier to label tense on Chinese documents. Ye et al. (2007) further reported that syntactic features contribute most to the marking of aspectual information. Liu et al. (2011) proposed a parallel mapping method to automatically generate annotated data. In particular, they used English verbs to label tense information for Chinese verbs via a parallel Chinese-English corpus.

It is reasonable to label such source-side verb to supervise the translation process since the tense of English sentence is often determined by verbs. The problem is that due to the diversity of English verb inflection, it is difficult to map such Chinese tense information into the English text. To our best knowledge, although above works attempt to serve for SMT, all of them fail to address how to integrate them into a SMT system.

## 2.2 Machine Translation with Tense

Dorr (1992) described a two-level knowledge representation model based on Lexical Conceptual Structures (LCS) for machine translation which integrates the aspectual information and the lexical-semantic information. Her system is based on an inter-lingual model and does not belong to a SMT system.

Olsen et al. (2001) relied on LCS to generate appropriately-tensed English translations for Chinese. In particular, they addressed tense reconstruction on a binary taxonomy (present and past) for Chinese text and reported that incorporating lexical aspect features of telicity can obtain a 20% to 35% boost in accuracy on tense realization.

Ye et al. (2006) showed that incorporating latent features into tense classifiers can boost the performance. They reported the tense resolution results based on the best-ranked translation text produced by a SMT system. However, they did not report the variation of translation performance after introducing tense information.

## 3 Preprocessing for Tense Modeling

In this paper, tense modeling is done on the target-side language. Since our experiments are done on Chinese to English SMT, our tense models are learned only from the English text. In the literature, the taxonomy of English tenses typically includes three basic tenses (i.e. present, past and future) plus their combination with the progressive and perfective aspects. Here, we consider four basic tenses: present, past, future and UNK (unknown) and ignore the aspectual information. Furthermore, we assume that one sentence has only one main tense but maybe has many subordinate tenses.

This section describes the preprocessing work of building tense models, which mainly involves how to extract tense sequence via tense verbs.

### 3.1 Tense Verbs

Lapata et al.(2006) used syntactic parse trees to find clauses connected with special aspect markers and collected them to train some special classifiers for temporal inference. Inspired by their work, we use the Stanford parser<sup>3</sup> to parse tense sequence for each sentence.

Take the following three typical sentences as examples, (a) is a simple sentence which contains two coordinate verbs, while (b) and (c) are compound sentences and (b) contains a quoted text.

(a) Japan's constitution renounces the right to go to war and prohibits the nation from having military forces except for self-defense.

(b) "We also hope Hong Kong will not be affected by diseases like the severe acute respiratory syndrome again!" , added Ms. Yang.

(c) Cheng said he felt at home in Hong Kong and he sincerely wished Hong Kong more peaceful and more prosperous.

Figure 1 shows the parse tree with Penn Treebank style for each sentence, which has strict level structures and POS tags for all the terminal words. Here, the level structures mainly contribute to extract main tense for each sentence (to be described in Section 3.2) and POS tags are utilized to detect *tense verbs*, i.e. verbs with tense information.

Normally, POS tags in the parse tree can distinguish five different forms of verbs: the base form (tagged as VB), and forms with overt endings D for

<sup>3</sup><http://nlp.stanford.edu/software/lex-parser.shtml>

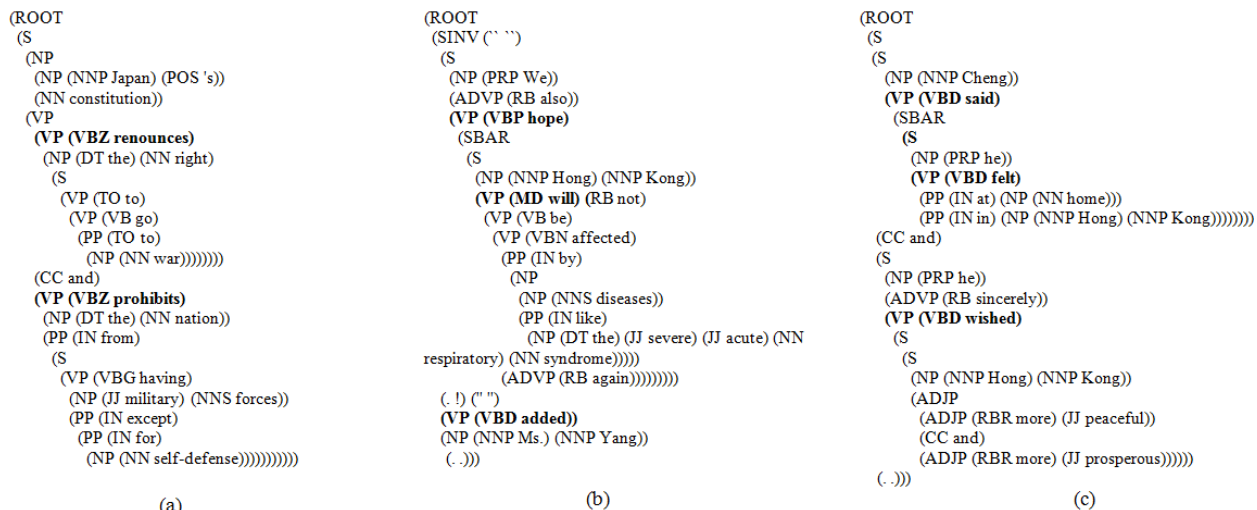


Figure 1: The Stanford parse trees with Penn Treebank style

past tense, G for present participle, N for past participle, and Z for third person singular present. It is worth noting that VB, VBG and VBN cannot determine the specific tenses by themselves. In addition, the verbs with POS tag “MD” need to be specially considered to distinguish future tense from other tenses.

Algorithm 1 illustrates how to determine what tense a node has. If the return value is not “UNK”, the node belongs to a *tense verb*.

---

**Algorithm 1** Determine the tense of a node.

---

**Input:**

The *TreeNode* of one parse tree, *leafnode*;

**Output:**

The tense, *tense*;

- 1:  $tense = \text{“UNK”}$
  - 2: Obtaining the POS tag  $lpostag$  from *leafnode*;
  - 3: Obtaining the word  $lword$  from *leafnode*;
  - 4: **if** ( $lpostag$  in [ $\text{“VBP”}$ ,  $\text{“VBZ”}$ ]) **then**
  - 5:      $tense = \text{“present”}$
  - 6: **else if** ( $lpostag == \text{“VBD”}$ ) **then**
  - 7:      $tense = \text{“past”}$
  - 8: **else if** ( $lpostag == \text{“MD”}$ ) **then**
  - 9:     **if** ( $lword$  in [ $\text{“will”}$ ,  $\text{“ll”}$ ,  $\text{“shall”}$ ]) **then**
  - 10:          $tense = \text{“future”}$
  - 11:     **else if** ( $lword$  in [ $\text{“would”}$ ,  $\text{“could”}$ ]) **then**
  - 12:          $tense = \text{“past”}$
  - 13:     **else**
  - 14:          $tense = \text{“present”}$
  - 15:     **end if**
  - 16: **end if**
  - 17: **return**  $tense$ ;
- 

### 3.2 Tense Extraction Based on Tense Verbs

As described in Section 1, the inter-tense (document-level) refers to the main tense of one sentence and the intra-tense (sentence-level) corresponds to all tense sequence of one sentence. This section introduces how to recognize the main tense and extract all useful tense sequence for each sentence.

The idea of determining the main tense is to find the *tense verb* located in the top level of a parse tree. According to the Penn Treebank style, the method to determine the main tense can be described as follows:

- (1) Traverse the parse tree top-down until a tree node containing more than one child is identified, denoted as  $S_m$ .
- (2) Consider each child of  $S_m$  with tag “VP”, recursively traverse such “VP” node to find a tense verb. If found, use it as the main tense and return the tense; if not, go to step (3).
- (3) Consider each child of  $S_m$  with tag “S”, which actually corresponds to subordinate clause of this sentence. Starting from the first subordinate clause, apply the similar policy of step (2) to find the tense verb. If not found, search remaining subordinate clauses.
- (4) If no tense verb found, return “UNK” as the main tense.

Here, “VP” nodes dominated by  $S_m$  **directly** are preferred over those located in subordinate clauses. This is to ensure that the main tense is decided by the top-level tense verb.

Take Figure 1 as an example, the main tense of sentence (a) and (b) can be determined only by step (2). The tense verb of “(VBZ renounces)” dominated in the “VP” tag determines that (a) is in present tense. Similarly the node “(VBD added)” indicates that (b) is in past tense. Sentence (c) needs to be further treated by step (3) since there is no “VP” nodes dominated by  $S_m$  directly. The node “(VBD said)” located in the first subordinate clause shows its main tense is “past”.

The next task is to extract the tense sequence for each sentence. They are determined by all tense verbs in this sentence according to the strict top-down order. For example, the tense sequence of sentence (a), (b) and (c) are {present, present}, {present, future, past} and {past, past, past}. In order to explore whether the main tense of intra-tense model has an impact on SMT or not, we introduce a special marker “\*” to denote the main tense. So the tense sequence marked with main tense of (a), (b) and (c) are {present\*, present}, {present, future, past\*} and {past\*, past, past}. It is worth noting, the intra-tense model (see Section 4) based on the latter tense sequence is different to the former.

## 4 N-gram-based Tense Models

### 4.1 Tense N-gram Estimation

After applying the previous method to extract tense for an English text corpus, we can obtain a big tense corpus.

Given the current tense is indexed as  $t_i$ , we call the previous  $n - 1$  tenses plus the current tense as tense n-gram.

Based on the tense corpus, tense n-gram statistics can be done according to the Formula 1.

$$P(t_i|t_{(i-(n-1))}, \dots, t_{(i-1)}) = \frac{\text{count}(t_{(i-(n-1))}, \dots, t_{(i-1)}, t_i)}{\text{count}(t_{(i-(n-1))}, \dots, t_{(i-1)})} \quad (1)$$

Here, the function of “count” return the tense n-gram frequency. In order to avoid doing specific smoothing work, we estimate tense n-gram probability using SRI language modeling (SRILM) tool (Stolcke, 2002).

To compute the probability of intra-tense n-gram, we first extract all tense sequence for each sentence

and put them into a new file. Based on this new file, we can get the intra-tense n-gram model via SRILM tool.

To compute the probability of inter-tense n-gram, we need to extract the main tense for each sentence at first. Then, for each document, we re-organized the main tenses of all sentences into a special line. After putting all these special lines into a new file, we can use SRILM to obtain the inter-tense n-gram model.

### 4.2 Characteristic of Tense N-gram Models

We construct n-gram-based tense models on English Gigaword corpus (LDC2003T05). This corpus is used to build language model for most SMT systems. It includes 30221 documents (we remove such files: file size is less than 1K or the number of continuous “UNK” tenses is greater than 5).

Figure 2 shows the unigram and bigram probabilities (**Log10-style**) for intra-tense and inter-tense.

The part (a) and (c) in Figure 2 refer to unigram. The horizontal axis indicates tense type, and the vertical axis shows its probabilities. The parts (a) and (c) also indicate “present” and “past” are two main tense types in news domain.

The part (b) and (d) refer to bigram. The horizontal axis indicates history tense. Each different colorful bar indicates one current tense. The vertical axis shows the transfer possibilities from a history tense to a current tense.

The part (b)<sup>4</sup> reflects transfer possibilities of tense types in one sentence. It also slightly reflects some linguistic information. For example, in one sentence, the probability of co-occurrence of “present → present”, “past → past” and “future → present” is more than other combinations, which can be against tense inconsistency errors described in Observation (1) and (2) (see Section 1). However, it seems strange that “present → past” exceeds “present → future”. We checked our corpus and found a lot of sentences like this—“the bill has been . . . , he said. ”.

The part (d) shows tense type can be switched between two neighbored sentences. However, it shows the strong tendency to use the same tense type for

<sup>4</sup>The co-occurrence of the “UNK” tense and other tense types in one sentence cannot happen, so the “UNK” tense is omitted.

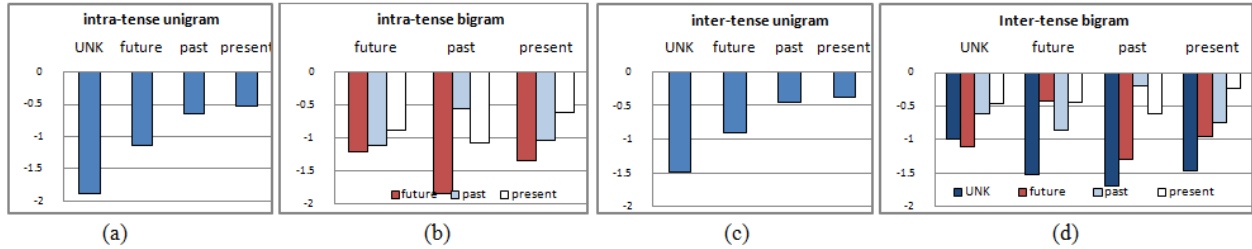


Figure 2: statistics of intra-tense and inter-tense N-gram

neighbored sentences. This statistics conform to the previous **observation (3)** very much.

## 5 Integrating N-gram-based Tense Models into SMT

In this section, we discuss how to integrate the previous tense models into a SMT system.

### 5.1 Basic phrase-based SMT system

It is well known that the translation process of SMT can be modeled as obtaining the best translation  $e$  of the source sentence  $f$  by maximizing following posterior probability (Brown et al., 1993):

$$\begin{aligned} e_{best} &= \arg \max_e P(e|f) \\ &= \arg \max_e P(f|e)P_{lm}(e) \end{aligned} \quad (2)$$

where  $P(e|f)$  is a translation model and  $P_{lm}$  is a language model.

Our baseline is a modified Moses, which 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 for obtaining the best translation according to the formula 3:

$$e_{best} = \arg \max_e \sum_{m=1}^M \lambda_m h_m(e, f) \quad (3)$$

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

### 5.2 The Workflow of Our System

Our system works as follows:

When a hypothesis has covered all source-side words during the decoding procedure, the decoder

first obtains tense sequence for such hypothesis and computes intra-tense feature  $F_s$  (see Section 5.3). At the same time, it recognizes the main tense of this hypothesis and associate the main tense of previous sentence to compute inter-tense feature  $F_m$  (see Section 5.3).

Next, the decoder uses such two additional feature values to re-score this hypothesis automatically and choose one hypothesis with highest score as the final translation.

After translating one sentence, the decoder caches its main tense and pass it to the next sentence. When one document has been processed, the decoder cleans this cache.

In order to successfully implement above workflow, we should firstly design some related features, then resolve another key problem of determining tense (especially main tense) for SMT output. They are described in Section 5.3 and 5.4 respectively.

### 5.3 Two Additional Features

Although the previous tense models show strong tendency to use the consistent tenses for one sentence or one document, other tense combinations also can be permitted. So we should use such models in a soft and dynamic way. We design two features: inter-tense feature and intra-tense feature. And the weight of each feature is tuned by the MERT script in Moses packages.

Given main tense sequence of one document  $t_1, \dots, t_m$ , the inter-tense feature  $F_m$  is calculated according to the following formula:

$$F_m = \prod_{i=2}^m P(t_i | t_{i-(n-1)}, \dots, t_{i-1}) \quad (4)$$

The  $P(\cdot)$  of formula 4 can be estimated by the formula 1. It is worth noting the first sentence of one

document often scares tense information since it corresponds to the title at most cases. To the first sentence, the  $P(\cdot)$  value is set to  $\frac{1}{4}$  (4 tense types).

Given tense sequence of one sentence  $s_1, \dots, s_e$  ( $e > 1$ ), the intra-tense feature  $F_s$  is calculated as follows:

$$F_s = \sqrt[e-1]{\prod_{i=2}^e P(s_i | s_{i-(n-1)}, \dots, s_{(i-1)})} \quad (5)$$

Here the square-root operator is used to avoid punishing translations with long tense sequence. It is worth noting if the sentence only contains one tense, the  $P(\cdot)$  value of formula 5 is also set to  $\frac{1}{4}$ .

Since the average length of intra-tense sequence is about 2.5, we mainly consider intra-tense bigram model and thus  $n$  equals to 2.<sup>5</sup>

#### 5.4 Determining Tense For SMT Output

The current SMT systems often produce odd translations partly because of abnormal word ordering and uncompleted text etc. For these abnormal translated texts, the syntactic parser cannot work well in our initial experiments, so the previous method to parse main tense and tense sequence of regular texts cannot be applied here too.

Fortunately, the solely utilization of Stanford POS tagger for our SMT output is not bad although it has the same issues described in Och et al. (2002). The reason may be that phrase-based SMT contains short contexts that POS tagger can utilize while the syntax parser fails.

Once obtaining a completed hypothesis, the decoder will pass it to the Stanford POS tagger and according to *tense verbs* to get all tense sequence for this hypothesis. However, since the POS tagger can not return the information about level structures, the decoder cannot recognize the main tense from such tense sequence.

Liu et al. (2011) once used target-side verbs to label tense of source-side verbs. It is natural to consider whether Chinese verbs can provide similar clues in an opposite direction.

Since Chinese verbs have good correlation with English verbs (described in section 6.2), we obtain

<sup>5</sup>In our experiment, the intra-tense bigram model can obtain the comparable performance to the trigram model. And the inter-tense trigram model can not exceed the bigram one.

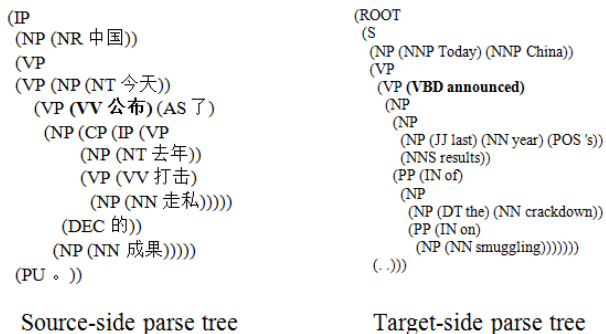


Figure 3: trees for parallel sentences

main tense for SMT output according to such *tense verb*, which corresponds to the “VV”(Chinese POS labels are different to English ones, “VV” refers to Chinese verb) node in the top level of the source-side parse tree. Take Figure 3 as an example, the English node “(VBD announced)” is a *tense verb* which can tell the main tense for this English sentence. The Chinese verb “(VV公布)” in the top-level of the Chinese parse tree is just the corresponding part for this English verb.

So, before translating one sentence, the decoder first parses it and records the location of one Chinese “VV” node which located in the top-level, denotes this location as  $S_{area}$ .

Once a completed hypothesis is generated, according to the phrase alignment information, the decoder can map  $S_{area}$  into the English location  $T_{area}$  and obtain the main tense according to the POS tags in  $T_{area}$ .

If  $T_{area}$  does not contain *tense verb*, such as the verb POS tags in the list of {VB, VBN, VBG}, which cannot tell tense type by themselves, our system permits to find main tense in the left/right 3 words neighbored to  $T_{area}$ . And the proportion that the top-level verb of Chinese has a verb correspondence in English can reach to 83% in this way.

## 6 Experimentation

### 6.1 Experimental Setting for SMT

In our experiment, SRI language modeling toolkit was used to train a 5-Gram general language model on the Xinhua portion of the Gigaword corpus. Word alignment was performed on the training parallel corpus using GIZA++ (Och and Ney, 2000) in two directions. 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. To see whether an improvement is statistically significant, we also conduct significance tests using the paired bootstrap approach (Koehn, 2004). In this paper, “\*\*\*” and “\*\*” denote p-values equal to 0.05, and bigger than 0.05, which mean significantly better, moderately better respectively.

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

Table 1: Corpus statistics

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 1 shows the statistics of these data sets (with document boundaries annotated).

## 6.2 The Correlation of Chinese Verbs and English Verbs

In this section, an additional experiment is designed to show English Verbs have close correspondence with Chinese Verbs.

We use the Stanford POS tagger to tag the Chinese and English sentences in our training corpus respectively at first. Then we utilize Giza++ to build alignment for these special Word-POS pairs. According to the alignment results, we find the corresponding relation for some special POS tags in two languages.

Chinese Verb POS	English POS	Number	
VV	Verb POS	VBD	89830
		VBP	27276
		VBZ	32588
		MD	40378
		VBG	86025
		VBN	75019
		VB	153596
	In sum:	504712	
	Other Non-Verb	149318	
Verb Corresponding Ratio		0.77169	

Table 2: The Chinese and English Verb Pos Alignment

The “Number” column of Table 2 shows the numbers of Chinese words with “VV” tag corresponding to English words with different verb POS tags.

We found Chinese verbs have more than 77% possibilities to align to English verbs in total. However, our method will fail when some special Chinese sentences only contain noun predicates.

## 6.3 Experimental Results

All the experiment results are showed on the table 3. Our Baseline is a modified Moses. The major modification is input and output module in order to translate using document as unit. The performance of our baseline exceeds the baseline reported by Gong et al. (2011) about 2 percent based on the similar training and test corpus.

System	BLEU Dev(%)	BLEU on Test(%)	
		BLEU	NIST
Moses_Md(Baseline)	29.21	28.30	8.4528
Baseline+ $F_m$	30.56	28.87(***)	8.7935
Baseline+ $F_s$	31.28	28.61(**)	8.5645
Baseline+ $F_s$ (*)	31.04	28.74(**)	8.6271
Baseline+ $F_m+F_s$	31.75	28.88(***)	8.7987
Baseline+ $F_m+F_s$ (*)	31.42	28.92(***)	8.8201

Table 3: The performance of using different feature combinations

The system denoted as “Baseline+ $F_m$ ” integrates the inter-tense feature. The performance boosts 0.57(\*\*\*) in BLEU score.

The system denoted as “Baseline+ $F_s$ ” integrates the intra-tense feature into the baseline. The improvement is less than the inter-tense model, only 0.31(\*\*). It seems the tenses in one sentence has more flexible formats than the document-level ones. It is worth noting, this method can gain higher performance on the develop data than the one of “Baseline+ $F_m$ ” while fail to improve the test data. Maybe the related weight is tuned over-fit.

The system denoted as “Baseline+ $F_s$ (\*)” is slightly different from “Baseline+ $F_s$ ”. This experiment is to check whether the main tense has an impact on intra-tense model or not (see Section 3.2). Here, the intra-tense model based on the tense sequence with main tense marker is slightly different to the model showed in Figure 2. The results are slightly better than the previous system by 0.13.

Finally, we use the two features together (Baseline+ $F_m+F_s$ ). The best way improved the performance by 0.62(\*\*\*) in BLEU score over our baseline.



## 6.4 Discussion

Table 4 shows special examples whose intra-tenses are changed in our proposed system. The example 1 and 2 show such modification can improve the BLEU score but the example 3 obtains negative impact. From these examples, we can see not only tense verbs have changed but also their surrounding words have subtle variation.

No.	BLEU	Translation sentence
1	8.64	Baseline: the gulf countries , the bahraini royal family members by the military career of part of the banned to their marriage stories like children , <b>have</b> become the theme of television films .
	19.71	Ours: the gulf country <b>is</b> a member of the bahraini royal family , a risk by military career <b>risks</b> part of the banned to their marriage like children , <b>has</b> become a story of the television and film industry .
2	17.16	Baseline:economists <b>said</b> that the sharp appreciation of the euro , in the recent investigation <b>continues</b> to have an impact on economic confidence , it <b>is</b> estimated that the economy is expected to rebound to pick up .
	24.25	Ours: economists <b>said</b> that the sharp appreciation of the euro , in the recent investigation <b>continued</b> to have an impact on economic confidence and therefore no reason to predict the economy expected to pick up a rebound .
3	73.03	Baseline: the middle east news agency <b>said</b> that , after the concerns of all parties concerned in the middle east peace process , israel and palestine , egypt , the united states , russia and several european countries <b>will</b> hold a meeting in washington .
	72.95	Ours: the middle east news agency <b>said</b> that after the concerns of all parties in the middle east peace process , israel and palestine , egypt , the united states , russia and several european countries <b>held</b> a meeting in washington .

Table 4: Examples with tense variation using intra-tense model

From the results showed on Table 3, the document-level tense model seems more effective than the sentence-level one. We manually choose and analyzed 5 documents with significant improvement in the test data. The part (a) of Figure 4 shows the main tense distributions of one reference. The main tense distributions for the baseline and our proposed system are showed in the part (b) and (c) respectively. These documents have different numbers of sentences but all less than 10. The vertical axis indicates different tense: 1 to “past”, 2 to “present”, 3 to “future” and 4 to “UNK”. It is obvious that our system has closer distributions to the ones of this reference.

The examples in Table 5 indicate the joint impact of inter-tense and intra-tense model on SMT. Sen-

### Src:

- 1)以色列屯民则封锁一条主要道路,抗议屯垦区遭迫击炮攻击。
- 2)巴勒斯坦解放组织领袖阿巴斯也获准前往约旦河西岸城镇伯利恒,这是过去四年来以色列当局首次准许巴勒斯坦重要领导人物参加圣诞节庆典。

### Ref:

- 1)Israeli settlers **blockaded** a major road to protest a mortar attack on the settlement area.
- 2)PLO leader Abbas **had** also been allowed to go to the West Bank town of Bethlehem , which **is** the first time in the past four years Israeli authorities **have** allowed a senior Palestinian leader to attend Christmas celebrations.

### Baseline:

- 1)israel **has** imposed a main road to protest by mortars attack .
- 2)the palestinian leader also **visited** the west bank cities and towns to bethlehem , which in the past four years , the israeli authorities **allowed** the palestinian leading figures attended the ceremony .

### Ours:

- 1)israel **has** imposed a main road to protest against the mortars attack .
- 2)leader of the palestinian liberation organization **have** also been allowed to go to the west bank towns , bethlehem in the past four years . this **is** the first time the israeli authorities **allow** palestinian leading figures attended the ceremony .

Table 5: the joint impact of inter- tense and intra-tense models on SMT

tence 1) and 2) are two neighbored sentences in one document. Both the reference and our output tend to use the same main tense type, but the former is in “past” tense and the latter is in “present” tense. The baseline cannot show such tendency. Although our main tense is different to the reference one, the consistent tenses in document level bring better translation results than the baseline. And the tenses in sentence level also show better consistency than the baseline.

## 7 Conclusion

This paper explores document-level SMT from the tense perspective. In particular, we focus on how to build document-level and sentence-level tense models and how to integrate such models into a popular SMT system.

Compared with the inter-tense model which greatly improves the performance of SMT, the intra-tense model needs to be further explored. The reasons are many folds, e.g. the failure to exclude quoted texts when modeling intra-tense, since tenses in quoted texts behave much diversely from normal texts. In the future work, we will focus on modeling intra-tense variation according to specific sentence types and using more features to improve it.

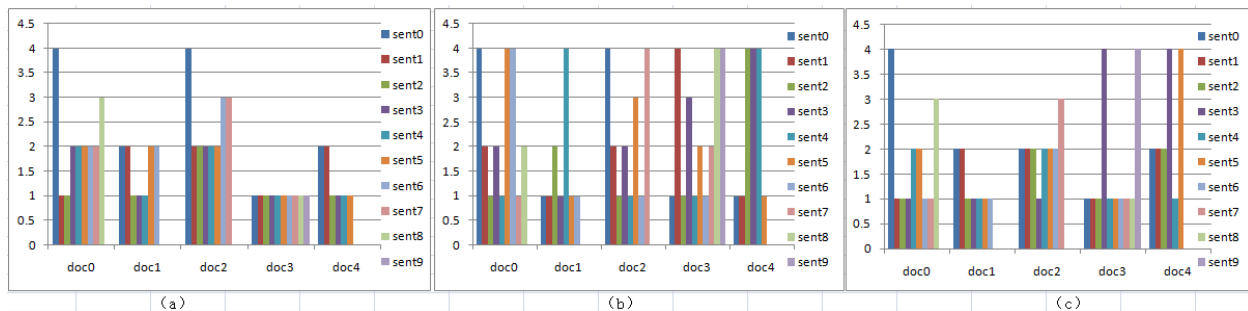


Figure 4: the comparison of the inter-tense distributions for reference, baseline and our proposed system

## Acknowledgments

This research is supported by part by NUS FRC Grant R252-000-452-112, the National Natural Science Foundation of China under grant No.90920004 and 61003155, the National High Technology Research and Development Program of China (863 Program) under grant No.2012AA011102.

## References

- P.F. Brown, S.A. Della Pietra, V.J. Della Pietra and R.L. Mercer. 1992. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263-311.
- Vilar David, Jia Xu, DĤaro L. F., and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 697-702, Genoa, Italy.
- Bonnie J. Dorr. 1992. A parameterized approach to integrating aspect with lexical-semantics for machine translation. In *Proceedings of of ACL-2002*, pages 257-264.
- Bonnie J. Dorr and Terry Gaasterland. 2002. Constraints on the Generation of Tense, Aspect, and Connecting Words from Temporal Expressions. Technical Reports from UMIACS.
- Zhengxian Gong, Min Zhang and Guodong Zhou. 2011. Cached-based Document-level Statistical Machine Translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 909-919.
- Philipp Koehn, Franz Josef Och ,and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of NAACL-2003*, pages 48-54.
- Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388-395.
- Mirella Lapata, Alex Lascarides. 2006. Learning Sentence-internal Temporal Relations. *Journal of Artificial Intelligence Research*, 27:85-117.
- Feifan Liu, Fei Liu and Yang Liu. 2011. Learning from Chinese-English Parallel Data for Chinese Tense Prediction. In *Proceedings of IJCNLP-2011*, pages 1116-1124, Chiang Mai, Thailand.
- Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of of ACL-2000*, pages 440-447.
- Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, et.al. 2002. A smorgasbord of Features for Statistical Machine Translation. In *Proceedings of NAACL-2004*, pages 440-447.
- Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL-2003*, pages 160-167, Sapporo, Japan, July.
- Mari Olsen, 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. In *Proceedings of MT Summit VIII*, Santiago de Compostella, Spain.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL-2002*, pages 311-318.
- Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901-904.
- Yang Ye and Zhu Zhang. 2005. Tense tagging for verbs in cross-lingual context: A case study. In *Proceedings of IJCNLP-2005*, pages 885-895.
- Yang Ye, V.li Fossum, Steven Abney. 2006. Latent features in Temporal Reference Translation. *Fifth SIGHAN Workshop on Chinese Language Processing*, pages 48-55.
- Yang Ye, Karl-Michael Schnelder, Steven Abney. 2007. Aspect Marker Generation in English-to-Chinese Machine Translation. *Proceedings of MT Summit XI*, pages 521-527, Copenhagen, Denmark.