

Disambiguating Tense, Aspect and Modality Markers for Correcting Machine Translation Errors

by

Anil Kumar Singh, Samar Husain, Harshit Surana, Jagadeesh Gorla, Chinnappa Guggilla, Dipti Misra
Sharma

in

*In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP). Borovets,
Bulgaria. 2007*

Report No: IIIT/TR/2007/75



Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
October 2007

Disambiguating Tense, Aspect and Modality Markers for Correcting Machine Translation Errors

Anil Kumar Singh, Samar Husain, Harshit Surana, Jagadeesh Gorla, Dipti Misra Sharma* and Chinnappa Guggilla†

Abstract

All languages mark tense, aspect and modality (TAM) in some way, but the markers don't have a one-to-one mapping across languages. Many errors in machine translation (MT) are due to wrong translation of TAM markers. Reducing them can improve the performance of an MT system. We used about 9000 sentence pairs from an English-Hindi parallel corpus. These were manually annotated with TAM markers and their mappings. Based on this corpus, we identify the factors responsible for ambiguity in translation. We present the results for learning TAM marker translation using CRF. We achieved an improvement of 17.88% over the baseline.

Keywords

Machine Translation, Tense, Aspect, Modality, TAM Markers

1 Introduction

Tense, aspect and modality are important elements of natural languages. They are needed for specifying the information about the world which is temporal in nature, or tell us something about the status of an action, or about the ability to perform an action. In some languages, they also govern the realization of a particular case marker. Different languages have different systems for marking such temporal (including aspectual and modal) information. In other words, TAM markers used by different languages don't have a one-to-one correspondence. TAM markers are not the only device used for expressing temporal information, but they can be very useful for NLP. They are a bit like function words. Like prepositions, even if not to the same extent, they can help in arriving at the correct syntactic and semantic analysis of a sentence. At the same time, they have an inherent meaning (even if ambiguous). This means that they are a bit like content words too. In this paper, we argue that these markers are under-utilized sources of linguistic information.

We first explain how we are defining TAM markers. Then we show that a significant percentage of errors in machine translation are due to wrong translation of TAM markers. We prepared an annotated parallel corpus to study the possibility of correcting these errors. The aim was to improve the performance of an

MT system. We annotated around 9000 sentence pairs from a sentence aligned English-Hindi parallel corpus with TAM markers and their mappings. Based on this corpus, we present the lists of most frequent markers and their translations. We also present the results of our experiments on learning translations of TAM markers using Condition Random Fields or CRF [4] and also show that we can improve the accuracy of an MT system by using this method. For our experiments, we have used the 0.73 version of the Shakti MT system [6].

1.1 What Exactly are TAM Markers

TAM markers are the combination of inflections (*en*, *ing*, *nA*¹, *tA*) and auxiliary verbs (*is*, *been*, *HE*, *thA*) or modals (*can*, *should*, *sakanA*, *paDZA*) or words indicating negativity (*not*, *naHIM*). These combinations together provide the information about tense, aspect and modality.

We can explain this by a hypothetical example from an English to Hindi machine translation system:

SL: So what happens now?

TL: to aba kyA HogA?

'So now what will-happen?'

TL (Default): * to aba kyA HonA HE?

In the example above, SL is the source language (English) sentence, TL is a correct translation in the target language (Hindi) followed by the literal English version of the correct Hindi translation. Finally, TL (Default) is the translation provided by the MT system, assuming that everything is correct except the TAM marker, because it was taken from a TAM dictionary with a one-to-one mapping.

In our terminology, we would say that in the SL sentence, *happens* has *PRES* (simple present) TAM marker, while the marker in TL for *HogA* is *gA* (future or hypothetical).

1.2 Empirical Evidence of the Problem

To get empirical evidence for our contention that wrong translation of TAM markers is a notable source of errors in MT, we extracted 250 random English sentences from the corpus. These sentences were run through the MT system. We manually checked these

*Language Technology Research Centre, IIT, Hyderabad, India, {aiklavya,samarhusain,surana.h}@gmail.com, jagadeesh.gorla@gmail.com and dipti@iiit.ac.in

†Applied Research Group, Satyam Computer Services Ltd, IISc campus, Bangalore, chinnappa.guggilla@satyam.com

¹ To represent text in Indian languages, we have used the RR notation. In this notation, capitalization roughly means longer length for vowels, and a small *h* after a consonant means aspiration.

English	PRES, PAST, to_0, ing, is_en, will_0, en, can_0, are_en, was_en, may_0, has_en, have_en, should_0, were_en, should_be_en, would_0, must_0, can_be_en, had_en, do_0, is_ing, to_be_en, by_ing,
Hindi	tA_HE, HE, nA, yA, thA, 0_sakatA_HE, gA, yA_jAtA_HE, yA_HE, nA.cAHiye, 0_raHA_HE, 0_kara, tA, yA_gayA_HE, yA_gayA, ye, tA_thA, yA_thA, 0_jAtA_HE, 0_gayA, HogA, yA_jAnA.cAHiye, yA_jA_sakatA_HE,

Table 1: Most frequent TAMs

sentences and marked the errors due to wrong translation of TAM markers. We found that 152 sentences had such errors. The number of wrongly translated TAM markers was 163 out of a total of 296 markers. This shows that there is empirical evidence of TAM markers being the cause of a significant number of errors in machine translation. If we can reduce these errors by using a better technique for TAM marker translation or for correcting such errors in the MT system output, we can improve the performance of the MT system.

2 Previous Work

Tense, aspect and modality have been studied extensively by linguists, both separately and as part of the study of temporal information encoded in natural languages. One of the most well known works in the first category is by Bybee et al. [2]. Their book discusses how tense, aspect and modality have evolved in different languages.

Vendler’s work [10] on verb classification with respect to time (or tense) is the basis of a lot of work on tense. In this work, he claimed that almost all the verbs can be classified into a few classes. Richenbach, in the classic work called ‘The Tenses of Verbs’ [8], suggests that the times of events can be located with respect to a deictic centre, which makes them similar to pronouns (the anaphoric view of tenses).

A lot of work has been done on temporal information from a computational point of view too. Dorr and Olsen [3] use a Lexical Conceptual Structure (LCS) based representation of Levin’s classes [5]. The aspectual classes are defined in terms of three features (telicity, dynamicity and durativity) and can be used to help in machine translation and generation.

Tense, aspect and modality in Indian languages have also been studied from a linguistic point of view. The book ‘Tense and Aspect in Indian Languages’ edited by Lakshmi Bai and Mukherji [1] contains a collection of a few such papers.

This paper has a different focus because we are concentrating on machine translation, whereas the focus of works mentioned above was language understanding or information extraction. As has been observed by many, some elements of machine translation may not require deep analysis of meaning.

In our opinion, TAM markers as a separate class of entities have not been given as much importance as they deserve, though they have been considered indirectly in the form of verb inflections and auxiliary verbs etc.

3 Problem Formulation

The problem we are addressing in this paper can be formulated as a disambiguation problem. In that sense it is similar to both preposition disambiguation and word sense disambiguation because TAM markers are like function words as well as content words, as mentioned earlier.

At a higher level of abstraction, the problem can also be formulated as a classification task. This formulation is more suitable than that of word sense disambiguation for our purposes because TAM markers form a closed class and the number of classes, though more than for prepositions, is small enough (50-200 for English and many Indian languages) to allow machine learning techniques such as CRF to be used.

If we classify TAM markers by considering contextual similarity of TAM markers, the problem becomes similar to POS tagging by using CRF:

$$t_i = f(s_i, c_i); \quad (1)$$

where s_i is the i^{th} TAM marker in the SL sentence and t_i is the translation of s_i , f represents a classification algorithm based on CRF, and c_i is the context for s_i . The CRF implementation that we have used was CRF++ [9]. The features we experimented with are described later in the Section-6.3.

4 Markers, Annotation and Dictionary

In this section we will first discuss the development of a set of TAM markers for a particular language, i.e., deciding on the set of TAM marker classes. We then discuss how a better TAM marker dictionary can be built. Finally, we describe how the parallel corpus was annotated with TAM markers and their mappings.

4.1 Developing a TAMMSet

How many TAM markers does a language have? Are they naturally and unambiguously very well defined? Not quite. We have to *design* a set of TAM markers for a particular language. This requires linguists, or at least well informed native speakers to sit down and list all possible TAM markers. The task of building this TAMMSet somewhat resembles the task of building a part of speech (POS) tagset for a particular language. In other words, even though they are linguistically significant, there may not be a universally acceptable set of markers for a language. Similarly, the set designed for one language may not be applicable for another.

English	Frequent Hindi Senses	English	Frequent Hindi Senses
PRES	HE, tA_HE, nA, yA_HE, gA, tA, 0_jAtA_HE, yA, ye, yA_jAtA_HE, 0_sakatA_HE, 0_kara, 0_raHA_HE,	PAST	yA, thA, tA_thA, HE, yA_thA, 0_gayA, nA, tA_HE, gA, yA_HE, 0_kara, yA_gayA, tA, tA_raHA
to_0	nA, ne_ke.liye, tA_HE, yA 0_sakatA_HE, HE, ye, 0_kara, gA, nA_HE,	ing	nA, tA_HE, 0_kara, HE, tA_HuA, 0_raHA_HE, tA, yA, ne_ke.liye,
is_en	yA_jAtA_HE, tA_HE HE, yA_gayA_HE, 0_kara	will_0	gA, HogA, tA_HE 0_sakatA_HE, HE

Table 2: Correspondence of some TAM categories (English-Hindi)

4.2 TAM Marker Dictionary

We had started with a basic TAM dictionary that was being used for machine translation. It was basic in the sense that it had only one-to-one mappings of TAM markers and the list of markers was shorter than the one we have after the new markers have been added during the annotation of the parallel corpus. Moreover, it was not compiled from the study of a corpus. Our aim was to build a proper TAM dictionary which can have one-to-many mappings corresponding to the possible senses of a TAM marker, just like an ordinary dictionary of words. The new dictionary was also to contain at least one example sentence in both SL and TL for every sense of a TAM marker.

4.3 TAM Marker Annotation

For annotating TAM markers, we selected at random more than 4000 short (up to 15 words) sentences and 5000 long sentences from a sentence aligned parallel English-Hindi corpus. These sentences were marked up using an interface by five different annotators. Some sets of sentences were then validated by a person different from the one who originally did the annotation. Annotators were given an initial list of SL and TL TAM markers, but they were asked to add a new marker if they thought it was required. An annotated sentence would look like this:

SL: So what [*happens*]_{PRES} now?

TL: to aba kyA [*HogA*]_{gA}?

Mapping: $PRES_1 \rightarrow gA_1$ (future)

4.4 Marker Lists from the Parallel Corpus

A list of most frequent markers (ranked according to frequency in the corpus) is given in Table-1. Table-2 gives a list of most frequent TAM marker mappings for English-Hindi. It is clear that the problem is not trivial and is a bit like word sense disambiguation.

5 Why TAM Markers: Another Example

There might be other ways of achieving the same kind of improvement in machine translation. Why use TAM markers? We will try to explain by an example how they can be useful. Consider the following text:

SL: We [*don't like*]_{PRES_not} that horse [*flying*]_{ing} in the sky. [*Shoot it down*]_{IMPER}.

TL: AsamAna meM [*uDZane vAlA*]_{ne_vAlA} vo ghoDZA HameM acchA [*naHIM laga raHA*]_{nahIM+0_rahA}. use [*mAra girAO*]_{0_0}.

‘sky in that-flies that horse we not-like. it shoot down.’

Mapping: $PRES_not_1 \rightarrow nahIM + 0_rahA_2, ing_2 \rightarrow ne_vAlA_1, IMPER_3 \rightarrow 0_O_3$

Now consider another variation of the same sentence pairs, *superficially* only slightly different:

SL: We [*don't like*]_{PRES_not} horses [*flying*]_{ing} in the sky. We [*shoot them down*]_{PRES}.

TL: AsamAna meM [*uDZane vAle*]_{ne_vAlA} ghoDZe HameM acche [*naHIM lagate*]_{nahIM+tA}. Hama unHeM [*mAra girAte HeM*]_{0_tA_HE}.

‘sky in that-fly horses we not-like. we them shoot down.’

Mapping: $PRES_not_1 \rightarrow nahIM + tA_2, ing_2 \rightarrow ne_vAlA_1, PRES_3 \rightarrow 0_tA_HE_3$

Note that in 0_0 , the first 0 is zero and is a place holder or wild card for verbs, while the second one is capital o , representing the inflection used for imperatives (*IMPER* for English).

What this example shows can be summarized as:

- The same TAM information can be expressed differently in different languages. TAM markers (at least partially) capture this difference. In the first set above, *PRES_not* (present with negation) gets translated as *nahIM+0_raHA*, while in the second as *nahIM+tA*. The only change in the SL sentence was that ‘that horse’ was substituted by ‘horses’. We can perform deep semantic analysis to get a correct translation in such a case, but it might not be possible in the near future for most (if not all) language pairs for obvious reasons. Or we could translate TAM marker separately.
- On the surface, only a slight change in the second case (‘we shoot’ instead of ‘shoot’) leaves the sentence no longer imperative, which changes the translation from (0_0 to 0_tA_HE). This will again be difficult to handle by semantic analysis, but is made easier if we use TAM markers.

6 Automatic Translation of TAM Markers

As indicated earlier, TAM markers can be translated by rule based, statistical or hybrid techniques. Theoretically, all these techniques can give good results. We have used a statistical or machine learning based technique.

6.1 Identifying TAM Markers

Though identifying TAM markers in the SL sentences is not the focus of the current work, one obvious method (which is already being used for machine translation) is through simple linguistic rules. In most cases it seems to work, provided that resources like dictionaries and morphological analyzer are available. However, for our experiments, we had the manually annotated markers in the corpus. We inserted them in the correct place in the MT system, so that we could see the results only for TAM marker translation, avoiding the errors in TAM marker identification.

6.2 Factors in TAM Marker Translation

The correct translation of an SL TAM marker depends on several factors. The preceding sections have already indicated them. In this section, we will take up all these factors and relate them to some possible solutions. The first item of information needed is, of course, the SL TAM marker. We assume that it is known since we are using TAM markup from the annotated corpus, rather than relying on the TAM marker computation module of the MT system, which can make mistakes (this is, of course, just for evaluation of TAM marker translation alone). A solution based only on the one-to-one TAM dictionary uses only this information. Three more factors are the distributions of SL and TL markers and their mappings in the corpus (or, ideally, in the language). A distributional similarity based solution, e.g. the IBM models [7] could take these factors into account. The most important factors for our proposed solution are the contexts in the SL sentence and in the TL output by the MT system (which we have to correct).

We have also tried to find the specific factors (or specific parts of the context) which determine this choice. One interesting example is given below:

SL: Who *wants*_{PRES} to lose their jobs?.

TL: apanI nOkarI kOna khonA *cAHegA*_{gA}?
'one's job who lose will-want?'

Mapping: $PRES_1 \rightarrow gA_1$

In this case (which is frequent in the corpus), the fact that the sentence is a question seems to determine that *PRES* will be translated as *gA* (future). Some other factors that seem to determine the choice of 'TAM sense' in the target language are the properties of the main verb, certain words (other than the verb), the type of the clause, etc. In fact, infinitives seem to have their own way of getting translated (see *to.0* in Table-2).

The proposed CRF based solution tries to take into account all the factors mentioned in this section. However, so far we have evaluated only with the context in the SL sentence, not with context in TL output by the MT system. This point is elaborated more in the next section.

6.3 Features for CRF

For now, we are only using the context from the SL sentence for learning and evaluation. We experimented on four sets of features for CRF. These were:

- **F1:** SL TAM marker, *verb_lex* (verb lexical item), *verb_cat* (verb category), *verb_lex-2* (word at a distance -2 from the current verb), *verb_lex-1*, *verb_cat-2*, *verb_cat-1*
- **F2:** SL TAM marker, *verb_lex*, *verb_cat*, *verb_lex-2/verb_lex-1* (combination of *verb_lex-2* and *verb_lex-1*), *verb_cat-2/verb_cat-1*
- **F3:** SL TAM marker, *verb_lex*, *verb_cat*, *verb_lex-2/verb_lex-1*, *verb_cat-2/verb_cat-1*, *head_lex-2/head_lex-1* (combination of lexical items for the head of the previous two chunks), *head_cat-2/head_cat-1*
- **F4:** SL TAM marker, *verb_lex*, *verb_cat*, 0 or 1 (1 if there is a conjunct except 'and' in the sentence, otherwise 0), *verb_lex-2/verb_lex-1*, *verb_cat-2/verb_cat-1*, *head_lex-2/head_lex-1*, *head_cat-2/head_cat-1*

7 Evaluation

In this section, we first describe the experimental setup and the evaluation method used by us. Then we present the results obtained.

7.1 Evaluation Method

For evaluation, we first conducted experiments on the four feature sets mentioned earlier to select the one which is likely to give to the best performance with CRF. We calculated the precision of TAM marker translation in three cases with the best feature set (F3):

- **A:** TAM dictionary with one-to-one mappings (baseline)
- **B:** MT system with output corrected using CRF (first evaluator)
- **C:** MT system with output corrected using CRF (second evaluator)

To prepare our training set we take the intermediate output from the MT system (after the POS tagger and the chunker) to get the context features. To evaluate which feature set was best (to be used for final evaluation), we divided the subset of the corpus into two parts (3470 sentences with 3908 markers, 530 sentences with 616 markers) by randomly selecting sentences. By training on the bigger set and testing on

Feature Set	Precision
F1	50.75%
F2	48.87%
F3	51.70%
F4	51.51%

Table 3: Results for four feature sets used for classification by CRF

	Precision
A. TAM Dictionary	46.05%
B. CRF-1	63.63%
C. CRF-2	64.22%

Table 4: Improvement in precision for TAM marker translation

the smaller one, we selected the most promising feature set. For this step, we used the markers in the parallel corpus as reference for evaluation. The feature set F3 gave the best performance (51.70% precision) and we used it for evaluation on the MT system.

For final evaluation, we took a subset of short sentences (6-15 words) from the annotated corpus and run them through the MT system and we correct the TAM marker in the MT system output based on learning by CRF. We tested on 439 sentences. The lower limit on the size of the sentences was to avoid fragments which were not really complete sentences. The higher limit was fixed because the MT system was not always able to process long sentences. Note that the output of the MT system was required to extract the features for learning. This is why we could not use the longer sentences.

7.2 Results

The evaluation was performed by two different evaluators. One was a professional translator while other was from computational linguistics background. Precision was calculated for default translation using the TAM dictionary (the baseline) and on the MT system with CRF corrected output.

The evaluators checked the TL markers in the context of their being meaningful keeping both the SL sentence and the intermediated MT system output in mind. The precision for the baseline was 46.05%. Learning by CRF gave a precision of 63.93%, which was significantly better than the baseline.

8 Observations Based on the Results

Based on the results obtained by correcting the MT system output using the CRF based marker classification, we present some observations about the errors and some suggestions for improving the results further:

1. Many of the errors were in translating *PRES* (simple present). On examining the training and testing data we found that this was because its distribution in the training and testing data was highly

imbalanced, i.e., our testing data was very unfair for evaluating the translation of this marker.

2. Another reason for errors in translating *PRES* is that it is more ambiguous. There are more ways in which it can be translated and many of them are quite frequent.
3. Some of the errors in translating *PRES* can be taken care of by simple rules. For example, in reported speech, the correct translation is usually *t_AHE*.

9 Conclusions and Future Work

We described the parallel corpus annotated with TAM markers and their mappings and listed the most frequent of them. We discussed why TAM markers are important for MT and gave linguistic and empirical evidence for this. The problem was formulated as a classification task. The technique we used for machine learning was CRF. We tested for four sets of features and selected the best one. Using this best feature set, we experimented on improving TAM translation. We were able to get a precision of 63.93%, which was significantly better than the baseline, which used a TAM dictionary with a one-to-one mapping. Based on our observation of the output, we suggested some ways to further improve the results. Another task for the future to use the context from the TL output given by the MT system, because TAM marker translation depends on the structure selected by the MT system for the translated sentence.

References

- [1] B. L. Bai and A. Mukherji, editors. *Tense and Aspect in Indian Languages*. Centre for Advanced Study in Linguistics, Osmania University, Hyderabad., 1993.
- [2] J. Bybee, R. Perkins, and W. Pagliuca. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. University of Chicago Press., 1994.
- [3] B. J. Dorr and M. B. Olsen. Deriving verbal and compositional lexical aspect for nlp applications. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics., 1997.
- [4] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.
- [5] B. Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago., 1993.
- [6] LTRC. A brief outline of shakti machine translation system, 2004. Language Technologies Research Centre, International Institute of Information Technology, Hyderabad, India, <http://ltrc.iiit.ac.in/showfile.php?filename=projects/shakti.php>.
- [7] V. J. D. P. Peter F. Brown, Stephen A. Della Pietra and R. L. Mercer. Mathematics of statistical machine translation: Parameter estimation. In *Computational Linguistics*, pages 19(2):263–311, 1993.
- [8] H. Reichenbach. The tenses of verbs. In *Elements of Symbolic Logic*. The Macmillan Company, New York., 1947.
- [9] S. Saravagi. Crf project page, 2005. A java implementation of Conditional Random Fields for sequential labeling. <http://crf.sourceforge.net/>.
- [10] Z. Vendler. Verbs and times. In *Linguistics in Philosophy*. Cornell University Press., 1967.