

Implicit Cues for Explicit Generation: Using Telicity as a Cue for Tense Structure in a Chinese to English MT System

Mari Olsen*

Microsoft Corporation
molsen@microsoft.com

Carol Van Ess-Dykema
U.S. Department of Defense
carol@umiacs.umd.edu

David Traum

USC/ICT
traum@cs.umd.edu

Amy Weinberg

University of Maryland
weinberg@umiacs.umd.edu

Abstract

In translating from Chinese to English, tense and other temporal information must be inferred from other grammatical and lexical cues. Tense information is crucial to providing accurate and fluent translations into English. Perfective and imperfective grammatical aspect markers can provide cues to temporal structure, but such information is optional in Chinese and is not present in the majority of sentences. We report on a project that assesses the relative contribution of the lexical aspect features of (a)telicity reflected in the Lexical Conceptual Structure of the input text, versus more overt aspectual and adverbial markers of tense, to suggest tense structure in the English translation of a Chinese newspaper corpus. Incorporating this information allows a 20% to 35% boost in the accuracy of the tense realization with a best accuracy rate of 92% on a corpus of Chinese articles.

1 Introduction

This paper advances the state of the art in lexicon design for MT by utilizing an interlingua where aspectual distinctions (telic versus atelic) that can be derived from verb classifications primarily influenced by considerations of argument structure can be used to fill lexical gaps in the source language that cannot be left unspecified in the target language. In translating from Chinese to English, tense must be inferred from other grammatical and lexical cues. For example, Chinese verbs do not necessarily specify whether the event described is prior or cotemporaneous with the moment of speaking. It is true that grammatical aspect information can be loosely associated with time, with imperfective aspect (Chinese 在 *zai-* and 着 *-zhe*) representing present time and perfective (Chinese 了 *le*) representing past time, (Chu, 1998; Li and Thompson, 1981). However, past-tense verbs do not need any aspect marking distinguishing them from present tense verbs. This is unlike English, which much more rigidly distin-

guishes past from present tense through use of suffixes. Thus, to generate an appropriate English sentence from its Chinese counterpart, we need to fill in a potentially unexpressed tense.

As an example, the final verb in sentence (1) is unmarked for aspect, but must be realized in the past tense.

- (1) 1965年前，我国
1965 year before, our_country
总共只有30万吨
altogether only have 30 ten_thousand ton
的造船能力，年产量是
de shipbuilding capacity, year output is
8万吨
8 ten_thousand ton

Before 1965 China had a total of only 300,000 tons of shipbuilding capacity and the annual output was 80,000 tons.

In our NLP applications, we use Lexical Conceptual Structures (LCS) (Jackendoff, 1983) as an interlingua, e.g., for machine translation. The primitives of the interlingua can capture both conceptual and syntactic generalizations among languages (Dorr et al., 1993).¹ Though LCS primitives deal with argument structures, (Dorr and Olsen, 1997a) have shown how to map the predicate types in the LCS to aspectual structure. Different predicate types, needed for argument structure mapping can encode whether an event is bounded in time (*telic*), or unbounded (*atelic*). We will rely on the lexical information of the verbs within a sentence to generate appropriately tensed English translations for Chinese.

2 Use of Aspect to Provide Temporal Information

We now discuss relevant aspectual features of sentences, and show how this can provide information

¹LCS representations in our system have been created for Korean, Spanish and Arabic, as well as for English and Chinese.

* Authors names are in alphabetical order

about the time of the situations presented in a sentence. Aspectual features can be divided into grammatical aspect, which is indicated by lexical or morphological markers in a sentence, and lexical aspect, which is inherent in the meanings of words.

2.1 Grammatical aspect

Grammatical aspect provides a viewpoint on situation (event or state) structure (Smith, 1997). Since imperfective aspect, such as the English PROGRESSIVE construction *be VERB-ing*, views a situation from within, it is often associated with present or contemporaneous time reference. On the other hand, perfective aspect, such as the English *have VERB-ed*, views a situation as a whole; it is therefore often associated with past time reference ((Comrie, 1976; Olsen, 1997; Smith, 1997) cf. (Chu, 1998)). The temporal relations are tendencies: although the perfective is found more frequently in past tenses (Comrie, 1976), both imperfective and perfective co-occur in some languages with past, present, and future tense. Grammatical aspect marking is optional in Chinese. This information can be marked by an optional post-verbal particle. When these particles are present, they provide helpful information and disambiguate the tense interpretation as shown in (2).

- (2) 1991年3月27日，
 1991 year 3 month 27 day US
 美国众议员沃尔夫等作为
 member_of Congress Wolfe class. as
 客人访问了北京第一监狱
 guest visit aspect Beijing de one prison
 On march 27,1991, Congressman Wolfe etc. visited Beijing Number one prison as guests.

Tense and/or aspect marking is required for English for both matrix and embedded clauses. For example, even a verb like *want*, which requires either a present infinitive or past oriented complement (and subject drop) (3), indicates whether the infinitive marks past or present time.

- (3) Wolfe wanted to {publicize / have publicized} the baseless criticism on various occasions.

Leaving out tense information, or getting it wrong during translation thus compromises both the fluency and the accuracy of the translation.

2.2 Adverbial Information

In addition to grammatical markers, certain adverbs place temporal restrictions on the tense of their associated clauses. For example 已 *yi*, and 已经 *yi jing* (already) imply the past tense, while 将 *jiang*,

将来 *jiang lai* (will), 会 *hui*, and 正在 *zheng zai* imply a future interpretation. When available, we want to use these cues to provide better translations

2.3 Lexical aspect

While grammatical aspect and overt temporal cues are clearly helpful in translation, there are many cases in our corpus in which such cues are not present, as in (4).

- (4) 特别 是从 1992 年到 1995
 especially is from 1992 year near_to 1995
 年， 外资 流入呈
 year, foreign-capital influx take-on
 直线上升趋势， 年均
 vertical soaring tendency, year average
 增长率为 19
 increase rate is 19
 Especially from 1992 to 1995, the foreign capital influx rose sharply. The average annual increase was 19 percent.

These are the hard cases, where we must infer tense or grammatical aspectual marking in the target language from a source that looks like it provides no overt cues. We will show however, that Chinese does provide implicit cues through its lexical aspect classes.

Lexical aspect refers to the type of situation denoted by the verb, alone or combined with other sentential constituents. The standard aspectual classes are based on three aspectual features: telicity, dynamicity and durativity. We focus on telicity, also known as BOUNDEDNESS. Verbs that are telic have an inherent end: *winning*, for example, ends with the finish line. Verbs that are atelic do not name their end: *running* could end with a distance *run a mile* or an endpoint *run to the store*, for example. Olsen (1997) proposed that aspectual interpretation be derived through monotonic composition of features as shown in Table 1. We focus on the telicity feature; the others do not concern us here.

According to many researchers, knowledge of lexical aspect—how verbs denote situations as developing or holding in time—correlates with the usual tense realization of verbs (Dowty, 1986; Moens and Steedman, 1988; Passoneau, 1988). In particular, Dowty suggests that, absent other cues, a telic event is interpreted as completed. Smith similarly suggests that in English all past events are interpreted as telic (Smith, 1997) (but cf. (Olsen, 1997)).

We note that these tendencies are heuristic, and not absolute. Nonetheless we will show first how we can read this information from the LCS, a representation not originally designed with this goal in mind,

Aspectual Class	Telic	Dynamic	Durative	Examples
State			+	<i>know, have</i>
Activity		+	+	<i>run, paint</i>
Accomplishment	+	+	+	<i>destroy</i>
Achievement	+	+		<i>notice, win</i>

Table 1: Lexical Aspect Features

and how this information can be used to guarantee better Chinese to English translation.

3 Aspect in Lexical Conceptual Structure

Our implementation of Lexical Conceptual Structure (LCS) — an augmented form of (Jackendoff, 1983; Jackendoff, 1990) — permits lexical aspect information to be computed from lexical entries for individual verbs as well as from composed representations for sentences, using uniform processes and representations. The LCS framework classifies verbs using primitives (GO, BE, STAY, etc.), types (Event, State, Path, etc.) and fields (Loc(ational), Temp(oral), Poss(essional), Ident(ificational), Perc(eptual), etc.). Our current working lexicon includes about 10,000 English verbs and 18,000 Chinese verbs. These verbs can be classified according to the primitives to derive aspectually related classes. Some examples of templates representing classes are shown in (5), along with an example of a verb in that class.

(5)

```
depart (go loc (* thing 2)
      (away_from loc (thing 2)
                   (at loc (thing 2)
                        (* thing 4)))
      (!!+ingly 26))

insert (cause (* thing 1)
       (go loc (* thing 2)
          ((* toward 5) loc (thing 2)
             ([at] loc (thing 2)
                  (thing 6))))
       (!!+ingly 26))
```

Telic verbs (and sentences) can be classed as either inherently telic or derived telic. Some verbs have an inherent endpoint, while others combine with other phrases to specify an end. Telic verbs constructed with paths will also have potential counterpart with an atelic verb plus prepositions or other lexical items to add the requisite path. *Depart*, for example, corresponds to *move away*, or something similar in another language.

We therefore identify telic sentences by the algorithm, formally specified in Figure 1 (simplified from (Dorr and Olsen, 1997b) [156]).

Given an LCS representation L:

1. Initialize: T(L):={∅T}
2. If Top node of L ∈ {CAUSE, LET, GO}
Then T(L):={+T}
3. If Top node of L ∈ {ACT, BE, STAY}
Then If Internal node of
L ∈ {TO, TOWARD, FOR_{Temp}}
Then T(L):={+T}
4. Return T(L)

Figure 1: Algorithm for LCS Telicity Determination

First the top node is examined for primitives that indicate telicity: if the top node is CAUSE, LET, GO, telicity is set to [+T], as with the verbs *break*, *destroy*, for example. If the top node is not a telic indicator (i.e., the verb is a basically atelic predicate such as *love* or *run*, telicity may still be indicated by the presence of complement nodes, e.g. a goal phrase (**to** primitive) in the case of *run*.

4 Predictions

Based on (Dowty, 1986) and others, as discussed above, we predict that sentences that Chinese sentences that lack grammatical aspect markers but have a telic LCS will better translate into English as the past tense, and those that lack telic identifiers will translate as present tense. Where present, grammatical aspect marking or adverbial marking can supersede the information provided by lexical aspect, with past adverbials and perfective markers yielding a past interpretation, and imperfective, and future oriented adverbials yielding present or future tense translations.

5 Implementation: a Chinese → English Machine Translation System

LSCes are used as the interlingua for our machine translation efforts. We have built a Chinese to English MT system focussed on translating newswire. Using the algorithm described below, the system aspectually types the relevant verbs using grammatical, or lexical aspect, or adverbial markers. The LCS thus provides the bridge from which the target-language sentence is generated.

Following the principles in (Dorr, 1993), lexical information and constraints on well-formed LCSes are used to *compose* an LCS for a complete sentence from a sentence parse in a source language. This composed LCS (CLCS) is then used as the starting points for generation into the target language, using lexical information and constraints for the target language.

The generation component consists of the following subcomponents:

Decomposition and lexical selection First, primitive LCSes for words in the target language are matched against CLCSes, and tree structures of covering words are selected. Ambiguity in the input and analysis represented in the CLCS is maintained (insofar as it is possible to realize particular readings using the target language lexicon), and new ambiguities are introduced when there are different ways of realizing a CLCS in the target language.

AMR Construction This tree structure is then translated into a representation using the Augmented Meaning Representation (AMR) syntax of instances and hierarchical relations (Langkilde and Knight, 1998a); however the relations include information present in the CLCS and LCSes for target language words, including theta roles, LCS type, and associated features.

Realization The AMR structure is then linearized, as described in (Dorr et al., 1998), and morphological realization is performed. The result is a lattice of possible realizations, representing both the preserved ambiguity from previous processing phases and multiple ways of linearizing the sentence.

Extraction The final stage uses a statistical extractor, using corpus-based bigram probabilities to pick an approximation of the most fluent realization (Langkilde and Knight, 1998b).

In order to realize sentences in English, we must have a tense feature, as discussed above. As a worst case, sentences with no tense feature could be given a random tense by the statistical extractor, or a default tense. Both of these options were tried, yielding poor results. For this reason, the realization algorithm has been augmented with the rules in (6), for creating a tense feature using other available information. Items prefixed with `:` indicate features present in the Verb AMR. `:aspect` refers to grammatical aspect from the analysis of the input. In this case, having the value `PERF` comes from (\pm *le*) being a direct subordinate of the verb. The `:telic` feature is computed using the algorithm in Figure 1. Finally, the `:headline` feature is assumed to be added

by a pre-processing phase, identifying an input sentence as being a newspaper headline.

- (6) If `:tense` feature in the input
 then use input value for `:tense`
 else if `:headline` +
 then `:tense` = `present`
 else if `:aspect` `PERF`
 then `:tense` = `past`
 else if adverb 已 or 已经
 then `:tense` = `past`
 else if adverb 会, 将, or 将来
 then `tense` = `present`
 else if `:telic` +
 then `:tense` = `past`
 else `:tense` = `present`

6 The Corpus

We have applied this machine translation system to a corpus of Chinese newspaper text from *Xinhua* and other sources, primarily in the economics domain. The genre is roughly comparable to the American *Wall Street Journal*. Chinese newspaper genre differs from other Chinese textual sources, in a number of ways, including:

- more complex sentence structure
- more extensive use of acronyms
- less use of Classical Chinese
- more representative grammar
- more constrained vocabulary (limited lexicon)
- abbreviations are used extensively in Chinese newspaper headlines

In order to test our hypothesis, we divided a 152 sentence newswire corpus into an 99 verb training set, and a 72 verb test set (some sentences had more than one main verb). The sentence structure is complex and stylized; with an average of 20 words per sentence in both the training and test corpora.

To evaluate the extent to which our predictions result in an improvement in translation, we have used a database of human translations of the sentences in our corpus as the ground truth, or gold standard. The translations were constructed to provide fluent English for comprehension, and not for the purposes of this experiment. In evaluating our results, we concentrate on how well the system did at matching past and present tenses to those provided by a human.

7 Results

The training corpus was used to refine the tense algorithm in (6), yielding success of greater than 90%. We have subsequently applied this algorithm to generate tense for the 72 additional clauses in the test

set, which had not been previously studied. Evaluation can be very difficult in a number of cases. Concerning tense, our “gold standard” is the set of human translations, previously constructed for these sentences. In many cases, there is nothing overt in the sentence which would specify tense, so a mismatch might not actually be “wrong”. Also, there are a number of sentences which were not directly applicable for comparison, such as when the human translator chose a different syntactic structure or a complex tense. These verbs either appeared in simple present, past, present or past perfect (has or had verb+ed), present or past imperfective (is verb+ing, was verb+ing) and their corresponding passive (is being kicked, was being kicked, have been kicked) forms. For cases like the present perfect (has kicked), we noted the intended meaning (e.g. past activity) expressed by the verb as well as the verb’s actual present perfective form. We scored the form as correct if the system translated a present perfective with past tense meaning as a simple past. The results of our evaluation are summarized in the tables below. The first table uses headline, grammatical aspect, adverbials, and lexical information. The results using only lexical information are summarized in the following table.

		generated tense	
		past	present
human translation	past	22	0
	present	6	44

Table 2: Preliminary Tense Results using all Info

		generated tense	
		past	present
human translation	past	18	4
	present	13	37

Table 3: Preliminary Tense Results using Lexical Only

Both results improve over our initial heuristic, which was to always use past tense (assuming this to be the default mode for newspaper article reporting). This heuristic yielded 57% accuracy on the training corpus, and considerably less for this test corpus. Using all information improved this to 92% correct, and even using just lexical information raised the accuracy to 76%. Review of our corpus leads to an interesting speculation about when grammatical aspect markers are selected. The present-oriented adverbs and aspect particles appeared exclusively on telic verbs. This suggests that an overt marker may be chosen if the lexical information would be misleading. This correlation did not hold of imperfectives, or past adverbials, but perhaps this is because in newswire, one must distinguish order of events

even when only past time is in question. For now, this must remain a speculation. Results are also clearly better than always picking present tense, or just using one of the features of grammatical or lexical aspect. We also note that in 2 cases of headlines, telicity alone would have predicted past tense, but the human translation used the present tense. Headlines are written using the historical present in English (“Man bites Dog”).

8 Conclusions

We therefore conclude that lexical and grammatical aspect can serve as a valuable heuristic for suggesting tense, in the absence of tense and other temporal markers. In addition, lexical aspect, as represented by the interlingual LCS structure, can serve as the foundation for language specific heuristics. Thus, the interlingual representation may be used to provide not only shared semantic and syntactic structure, but also the building blocks for language-specific heuristics for mismatches between languages. More importantly, it can be used to infer information not overtly present in the string or syntactic structure of a language, leading to fluent and accurate translation.

9 Future Research

There are a number of other directions we intend to pursue in extending this work. First, we plan to try this on larger scale corpora. We also plan to extend our work to uncovering implicit discourse relations, capitalizing on the insight that completed events usually indicate sequentiality while uncompleted events are co-temporaneous. We would also like to extend this approach to other information contained in the LCS (e.g., causality), and to investigate further whether we can predict when an overt marker is likely to be used.

References

- Chauncey C. Chu. 1998. *A Discourse Grammar of Mandarin Chinese*. Peter Lang Publishing, Inc., New York, NY.
- Bernard Comrie. 1976. *Aspect*. Cambridge University Press, Cambridge, MA.
- Bonnie J. Dorr and Mari Broman Olsen. 1997a. Aspectual Modifications to a LCS Database for NLP Applications. Technical Report LAMP TR 007, UMIACS TR 97-23, CS TR 3763, University of Maryland, College Park, MD.
- Bonnie J. Dorr and Mari Broman Olsen. 1997b. Deriving Verbal and Compositional Lexical Aspect for NLP Applications. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 151–158, Madrid, Spain, July 7-12.

- Bonnie J. Dorr, James Hendler, Scott Blanksteen, and Barrie Migdaloff. 1993. Use of Lexical Conceptual Structure for Intelligent Tutoring. Technical Report UMIACS TR 93-108, CS TR 3161, University of Maryland.
- Bonnie J. Dorr, Nizar Habash, and David Traum. 1998. A Thematic Hierarchy for Efficient Generation from Lexical-Conceptual Structure. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*, pages 333–343, Langhorne, PA, October 28-31.
- Bonnie J. Dorr. 1993. *Machine Translation: A View from the Lexicon*. The MIT Press, Cambridge, MA.
- David Dowty. 1986. The Effects of Aspectual Class on the Temporal Structure of Discourse: Semantics or Pragmatics? *Linguistics and Philosophy*, 9:37–61.
- Ray Jackendoff. 1983. *Semantics and Cognition*. The MIT Press, Cambridge, MA.
- Ray Jackendoff. 1990. *Semantic Structures*. The MIT Press, Cambridge, MA.
- Irene Langkilde and Kevin Knight. 1998a. Generation that Exploits Corpus-Based Statistical Knowledge. In *Proceedings of COLING-ACL '98*, pages 704–710.
- Irene Langkilde and Kevin Knight. 1998b. The Practical Value of N-Grams in Generation. In *International Natural Language Generation Workshop*.
- Charles Li and Sandra Thompson. 1981. *Mandarin Chinese: A functional reference grammar*. University of California Press, Berkeley, CA.
- Marc Moens and Mark Steedman. 1988. Temporal Ontology and Temporal Reference. *Computational Linguistics: Special Issue on Tense and Aspect*, 14(2):15–28.
- Mari Broman Olsen, Bonnie J. Dorr, and Scott C. Thomas. 1998. Enhancing Automatic Acquisition of Thematic Structure in a Large-Scale Lexicon for Mandarin Chinese. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*, pages 41–50, Langhorne, PA, October 28-31.
- Mari Broman Olsen. 1997. *A Semantic and Pragmatic Model of Lexical and Grammatical Aspect*. Garland, New York.
- Rebecca Passoneau. 1988. A Computational Model of the Semantics of Tense and Aspect. *Computational Linguistics: Special Issue on Tense and Aspect*, 14(2):44–60.
- Carlota Smith. 1997. *The parameter of aspect*. Kluwer, Dordrecht, 2nd edition.