# Propbanking in Parallel

**Paul Kingsbury, Nianwen Xue, Martha Palmer**
Department of Computer and Information Science
University of Pennsylvania
{kingsbur, xueniwen, mpalmer}@unagi.cis.upenn.edu

### Abstract

This paper describes an effort to provide semantic role annotation for parallel Chinese/English corpora that we believe has the potential of benefiting statistical machine translation. This level of annotation, called a Parallel Proposition Bank, abstracts away from divergences in word order and syntactic categories to facilitate a mapping from a clausal structure in one language to the corresponding clausal structure in the other language. It collects together split arguments, making it easier to find their foreign language counterparts. It also provides for a level of coarse-grained word sense disambiguation based primarily on differences in subcategorization frames that could simplify the task of lexical choice. Although there are still many language specific characteristics of the semantic annotation, it moves us one step closer to a general semantic representation that is language independent.

## Introduction

Concurrent with the completion of the PropBank project at Penn (Palmer et al (submitted), Kingsbury and Palmer 2002), the decision was made to extend the annotation methodology both to independent corpora in other languages and to multilingual parallel corpora. The intention is first to gain resources for shallow semantic analysis in languages other than English; thus, monolingual PropBanking efforts have begun for Chinese (Xue and Palmer 2003), Korean and one is planned for Arabic. A second, more salient goal is the facilitation of machine translation systems. Just as there is evidence that syntactic parses improve the accuracy of MT systems (Yamada and Knight 2001, Charniak etal. 2003), it is expected that semantic parses will also improve accuracy, by showing explicit dependency relationships between elements of a sentence. PropBanking further includes a degree of coarse-grained sense-tagging which could also facilitate accurate translations.

PropBanking in parallel requires a number of resources. A first obvious step is the collection or creation of a parallel corpus annotated with syntactic structures. The Penn Chinese Treebank comprises almost 250K words of Xinhua news and 250K words of Sinorama (a Taiwanese multilingual news magazine) (Xue et al. 2004). There is on-going effort at the University of Pennsylvania to treebank the English translation of the first 100 thousand words of the treebanked Xinhua news, as well as the corresponding 250K word English Sinorama corpus. More important is the pre-existence of argument-structure lexicons for each of the languages in question. More than 3300 lexical items of English already have entries in the Propbank frame lexicon, and there are more than 4500 Chinese PropBank entries. A third component of the parallel propbanking endeavor is to explore the transferability between the languages at the level of frameset. It is hoped that this transferability can be exploited in future Machine Translation systems.

## The Propbank

### Generalities and the English Propbank

PropBank is a shallow semantic parse of running text, marking the argument structure of the verbs and deverbal adjectives. It comprises two separate but interdependent parts. The first is an annotated corpus wherein every verb and its arguments are explicitly marked. The corpus in question for English is the Wall Street Journal portions of the Penn TreeBank II (Marcus et al, 1994), while for Chinese the corpus is the Chinese TreeBank (Xue et al, 2004). Of more interest is the second part of the PropBank resource, the so-called 'frames files.' These are collectively a lexicon detailing the specific arguments expected to appear with any given verb. Arguments are assigned a (relatively) theory-neutral numbered label and are assigned a verb-specific mnemonic label. Different senses of a verb are assigned to different 'framesets' containing independent definitions of arguments. Senses are defined on both semantic and syntactic grounds. For example, the English verb 'afford' is seen in contexts such as the following:

1. These days Nissan can afford that strategy, even
   though profits aren't exactly robust. (wsj_0286)
2. Last year the public was afforded a preview of Ms.
   Bartlett's creation in a tablemodel version, at a BPC
   exhibition. (wsj_0984)

Although each example shows two arguments, the passive morphology on the second sentence shows that a third argument must be possible, providing a syntactic motivation for the framing of 'afford' as follows:

afford.01 'be able to sustain the cost of something'
arg0: entity sustaining cost
arg1: costly thing

afford.02 'provide, make available'
arg0: provider
arg1: thing provided
arg2: recipient

Framesets are also distinguished when the meanings of the usages are sufficiently different, even if the number of roles is the same. For example, the verb 'stem' also takes two framesets[1], each with two roles, on the basis of sentences such as the following:

3. Travelers Corp.'s third-quarter net income rose 11%, even though claims stemming from Hurricane Hugo reduced results $40 million. (wsj_0144)
4. If the company can start to ship during this quarter, it could stem some, if not all of the red ink, he said. (wsj_1973)

Under most circumstances a relatively proficient speaker of English will be able to distinguish between these senses, motivating their classification into separate framesets.

stem.01 'arise'
arg1: entity arising, coming about
arg2: arising from what?

stem.02 'stanch, cause to stop flowing'
arg0: causer of non-flowing
arg1: thing no longer flowing

Verb senses are thus defined at a level considerably more coarse-grained than the senses used in WordNet (Palmer, et. al., 2004), but the disambiguation still results in an explosion of related verbs. The English TreeBank contains approximately 3300 separate lexical items identified as verbs, but even the coarse-grained distinctions produce more than 4600 framesets.

## Special Issues for the Chinese Propbank

The same annotation philosophy has been extended to the Penn Chinese Proposition Bank (Xue and Palmer, 2003). In Chinese, the same syntactic alternations that form the basis for the English PropBank annotation also exist in robust quantities, even though it may not be the case that the same exact verbs (meaning verbs that are close translations of one anther) have the exact same range of syntactic realization for Chinese and English. For example, in (5), "xin-nian/New Year zhao-dai-hui/reception" plays the same role in (a) and (b), even though it occurs in different syntactic positions. This regularity is captured by assigning the same argument label ARG1 to both instances. It is worth noting that the predicate "ju-xing/hold" does not have passive morphology in (5a), despite what its English translation suggests. Like the English Propbank, the adjunct-like elements receive more general labels like TMP or LOC. The tag set for Chinese and English PropBanks are to a large extent similar and more details can be found in (Xue and Palmer, 2003).

5. a. [$_{\text{ARG1}}$ *xin-nian*/New Year *zhao-dai-hui*/reception] [$_{\text{ARGM-TMP}}$ *jin-tian*/today][$_{\text{ARGM-LOC}}$ *zai*/at *diao-yu-tai*/Diaoyutai *guo-bin-guan*/state guest house] *ju-xing*/hold
"A New Year reception was held in Diaoyutai State Guest House today."

b. [$_{\text{ARG0}}$ *tang-jia-xuan*/Tang Jiaxuan] [$_{\text{ARGM-TMP}}$ *jin-tian*/today] [$_{\text{ARGM-LOC}}$ *zai*/at *diao-yu-tai*/Diaoyutai *guo-bin-guan*/state guest house] *ju-xing*/hold [$_{\text{ARG1}}$ *xin-nian*/New Year *zhao-dai-hui*/reception]
"Tang Jiaxuan was holding the New Year Reception in Diaoyutai State Guest House today."

For polysemous verbs we also distinguish different framesets. (6) and (7) illustrate the different framesets of "tong-guo/pass", which correspond with major senses of the verb, loosely defined. The frameset in (6) roughly means "pass by voting" while the frameset illustrated by (7) means "pass through".

6. a. [$_{\text{ARG0}}$ *mei-guo*/the U.S. *guo-hui*/Congress] *zui-jin*/recently *tong-guo*/pass *le*/ASP [$_{\text{ARG1}}$ *zhou-ji*/interstate *yin-hang-fa*/banking law]
"The U.S. Congress recently passed the inter-state banking law."
b. [$_{\text{ARG1}}$ *zhou-ji*/interstate *yin-hang-fa*/banking law] *zui-jin*/recently *tong-guo*/pass *le*/ASP
"The inter-state banking law passed recently."

7. a. [$_{\text{ARG0}}$ *huo-che*/train] *zhen-zai*/now *tong-guo*/pass [$_{\text{ARG1}}$ *sui-dao*/tunnel]
"The train is passing through the tunnel."
b. [$_{\text{ARG0}}$ *huo-che*/train] *zheng-zai*/now *gong-guo*/pass.
"The train is passing."

Despite these similarities between the languages, there are also some Chinese-specific issues that have to be dealt with in the process of creating frame files. One issue is the disambiguation of preverbal prepositional phrases. As illustrated in (8), these preverbal PPs can be dependent on the verb, as in (8a), or the postverbal NP as in (8b). In English, since all such PPs are postverbal, this disambiguation can be done straightforwardly in syntax by attaching them at different levels. Such a simple solution does not exist for Chinese. Instead, this is handled as part of the PropBanking effort by marking verb-dependent PPs, such as that of (8a), as a semantic argument of the verb. The noun-dependent PP in (8b) will be related to the post-verbal NP and will have no predicate-argument label relative to the verb.

8. a. *zai*/at *jiu-hui*/banquet *shang*/on *cai*/Cai *da-shi*/ambassador **[**$_{\text{PP}}$ *dui*/to *yi-xiang*/always *guan-xin*/support *zu-guo*/motherland *jian-she*/development *de*/DE *hai-wai*/overseas *tong-bao*/compatriot**]** [$_{\text{V}}$ *fa-biao*/deliver] *le*/ASP [$_{\text{NP}}$ *re-qing*/enthusiam *yang-yi*/overflow *de*/DE *jiang-hua*/speech].
"At the banquet, Ambassador Cai made an enthusiastic speech to the overseas compatriots."

b. *zeng-yin-quan*/Zeng Yinquan **[**$_{\text{PP}}$ *jiy*/on *jian-li*/establish *guo-ji*/international *jin-rong*/financial *xin*/new *zhi-xu*/order**]** [$_{\text{V}}$ *fa-biao*/express] [$_{\text{NP}}$ *jian-jie*/view] .

[1] This ignores two other possible senses of 'stem' which do not happen to occur in the corpus, namely 'reduce to just a stem' as in a morphological stemmer and 'remove the stems of something which inherently has a stem' as in stemmed cherries.

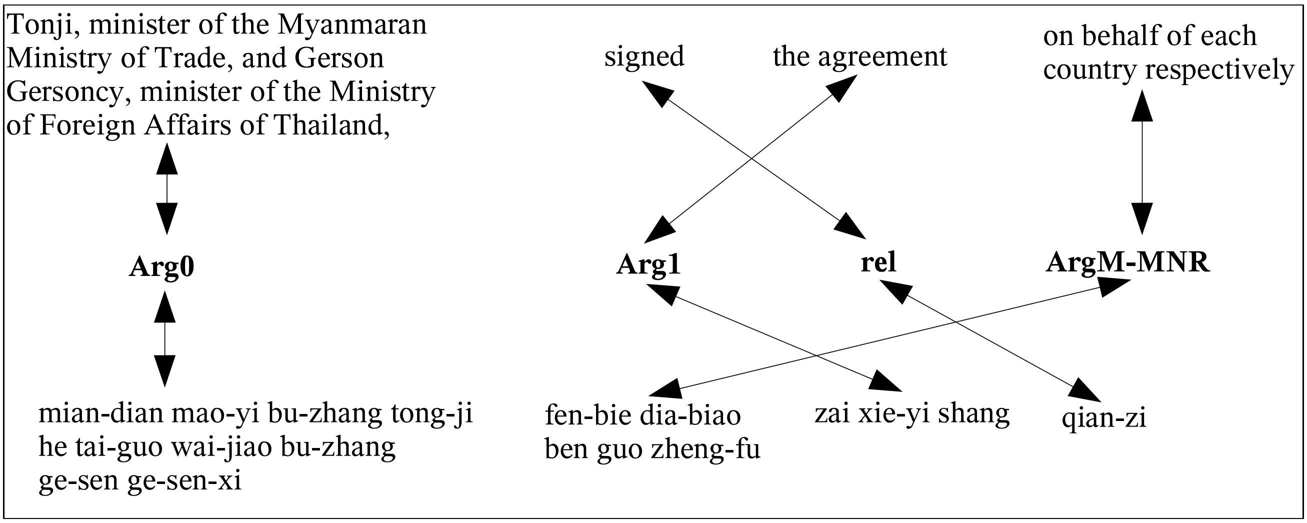
Tonji, minister of the Myanmaran Ministry of Trade, and Gerson Gersoncy, minister of the Ministry of Foreign Affairs of Thailand,
signed
the agreement
on behalf of each country respectively
Arg0
Arg1
rel
ArgM-MNR
mian-dian mao-yi bu-zhang tong-ji he tai-guo wai-jiao bu-zhang ge-sen ge-sen-xi
fen-bie dia-biao ben guo zheng-fu
zai xie-yi shang
qian-zi

Figure 1: Mapping between Chinese and English arguments

"Zeng Yinquan expressed his own view on the establishment of a new international financial order."

Another phenomenon which is much more common in Chinese than in English is split arguments. One such split is between the possessor (PSR) and the possessee (PSE). Here the possessor and possessee are abstract notions and do not necessarily indicate a strict possession relation. This is illustrated in (9).

9. [ARG1-psr *zhong-guo*/China *jing-ji*/economy *zeng-zhang*/growth] *ye*/also *jiang*/will [V *fang-man*/slow down] [ARG1-pse *su-du*/speed]
"The speed of Chinese economic growth will also slow down."

## Transferability of Framesets

The value of the PropBanking effort lies in the fact that the semantic representations implemented in the frame files of the two languages abstract away from the syntactic idiosyncrasies of the individual languages and create a platform where the predicate-argument structure mapping can take place. If these mappings can be recovered automatically, then it will have a profound impact on machine translation. Although the extent to which such mapping can be performed in a straightforward manner is yet to be determined, a preliminary examination shows that the PropBank annotations would facilitate such a mapping in a number of ways. First, the PropBank representation abstracts away from divergences in the word order and the syntactic category of the two languages and allows for a straightforward mapping at the predicate-argument structure level. This is illustrated in (10) and graphically in Figure 1.

10. [ARG0 Tonji, minister of the Myanmaran Ministry of Trade, and Gerson Gersoncy, minister of the Ministry of Foreign Affairs of Thailand], [FRAMESET.01 signed] [ARG1 the agreement] on behalf of each country respectively.
[ARG0 *mian-dian*/Myanmar *mao-yi*/trade *bu-zhang*/minister *tong-ji*/Tonji *he*/and *tai-guo*/Thailand *wai-jiao* *bu-zhang*/foreign minister *ge-sen ge-sen-xi*/Gerson Gersoncy] *fen-bie*/respectively *dai-biao*/represent *ben*/own *guo*/country *zheng-fu*/government [ARG1 *zai*/at *xie-yi*/agreement *shang*/above] [FRAMESET.01 *qian-zi*/sign].

Second, the PropBank annotation also abstracts away from the split argument phenomenon in the two languages. Split arguments may occur in different places and with different predicates in the two languages, but the PropBank annotation addresses this by marking the pieces as belonging to the same argument. This is illustrated in (11), adapted from (9):

11. [ARG1-psr *zhong-guo*/China *jing-ji*/economy *zeng-zhang*/growth] *ye*/also *jiang*/will [V *fang-man*/slow down] [ARG1-pse *su-du*/speed]
[ARG1-pse The speed] [ARG1-pse of Chinese economic growth] will also slow down.

Having the frameset information also enables us to map framesets that have compatible argument structures across languages. In many cases the framesets of a verb in one language map to different lexical items in another. For example, "leave" has two framesets and each takes a different set of arguments. They are mapped to different lexical items in Chinese:

leave.01: *li-kai*
Arg0: entity leaving
Arg1: place left
Arg2: attribute of Arg1

12. This flight leaves Shanghai at midnight.
*hang-ban*/flight *wu-ye*/midnight *li-kai*/leave *shang-hai*/Shanghai

leave.02: *liu-gei*
Arg0: giver
Arg1: thing given
Arg2: benefactor

13. John left Mary a big fortune.
*yue-han*/John *liu-gei*/leave *ma-li*/Mary *yi*/one *da-bi*/big sum *cai-chan*/fortune

## Conclusion

This paper has described the basis of the PropBank annotation that is being applied to parallel Chinese/English corpora. The English and Chinese PropBanks provide a level of annotation that highlights the dependency structure of a clause and the semantic roles played by the dependents. It abstracts away from surface idiosyncrasies such as word order, syntactic category and split constituents. The expectation is that this level of annotation, in addition to aiding the development of increasingly sophisticated monolingual information processing tools, will also prove useful to various kinds of machine translation systems. Transfer-based machine translation approaches could benefit from corpus-based transfer lexicons extracted from PropBanked parallel corpora. Statistical machine translation systems could re-rank potential target language outputs based on the similarity between their semantic role labels and those of the source language sentence. Although still preserving many language-specific characteristics, this level of annotation is one step closer to a general-purpose semantic representation.

## Acknowledgments

Work on the English PropBank has been funded by DARPA grant N66001-00-1-8915 and DOD grant MDA904-00-C-2136. The Chinese PropBank has been funded by DOD grant MDA904-02C0412. The Parallel PropBanks are being funded by NSF ITR Supplement 0343005.

## References

Charniak, E., Knight, K. & Yamada K. (2003) Syntax-based Language Models for Machine Translation. *Proceedings of MT Summit IX 2003*. New Orleans.

Kingsbury, P. & Palmer, M. (2002) From Treebank to Propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (LREC-2002), Las Palmas, Spain.

Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., & Schasberger, B. (1994) The Penn Treebank: Annotating predicate argument structure. In *ARPA Human Language Technology Workshop*, pp 114-119, Plainsboro NJ.

Ng, H.T., Wang, B., & Chan, Y.S. (2003). Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In the *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (ACL-03). Sapporo, Japan.

Palmer, M., Gildea, D. & Kingsbury, P. (submitted) The Proposition Bank: An Annotated Corpus of Semantic Roles, submitted to *Computational Linguistics.*

Palmer, M., Babko-Malaya, M., Dang, H., Different Sense Granularities for Different Applications, *2$^{nd}$ Workshop on Scalable Natural Language Understanding Systems, at HLT/NAACL-04,* Boston, Mass, May 6, 2004.

Xue, N. & Palmer, M. (2003) Annotating Propositions in the Penn Chinese Treebank. In *Proceedings of the Second Sighan Workshop,* in conjunction with ACL'03, Sapporo, Japan.

Xue, N., Xia, F., Chiou, F. & Palmer, M. 2004. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus, *Natural Language Engineering,* 10(4):1-30.

Yamada, K. & Knight, K. 2001. A Syntax-based Statistical Translation Model. *Proceedings of the Conference of the Association for Computational Linguistics,* (ACL-2001).