A Parallel Proposition Bank II for Chinese and English

Martha Palmer, Nianwen Xue, Olga Babko-Malaya, Jinying Chen, Benjamin Snyder
Department of Computer and Information Science
University of Pennsylvania
{mpalmer/xueniwen/malayao/Jinying/bsnyder3}@linc.cis.upenn.edu

Abstract

The Proposition Bank (PropBank) project is aimed at creating a corpus of text annotated with information about semantic propositions. The second phase of the project, PropBank II adds additional levels of semantic annotation which include eventuality variables, co-reference, coarse-grained sense tags, and discourse connectives. This paper presents the results of the parallel PropBank II project, which adds these richer layers of semantic annotation to the first 100K of the Chinese Treebank and its English translation. Our preliminary analysis supports the hypothesis that this additional annotation reconciles many of the surface differences between the two languages.

1 Introduction

There is a pressing need for a consensus on a task-oriented level of semantic representation that can enable the development of powerful new semantic analyzers in the same way that the Penn Treebank (Marcus et al., 1993) enabled the development of statistical syntactic parsers (Collins, 1999; Charniak, 2001). We believe that shallow semantics expressed as a dependency structure, i.e., predicate-argument structure, for verbs, participial modifiers, and nominalizations provides a feasible level of annotation that would be of great benefit. This annotation, coupled with word senses, minimal co-reference links, event identifiers, and discourse and temporal relations, could provide the foundation for a major advance in our ability to automatically extract salient relationships from text. This will in turn facilitate breakthroughs in message understanding, machine translation, fact retrieval, and information retrieval. The Proposition Bank project is a major step towards providing this type of annotation. It takes a practical approach to semantic representation, adding a layer of predicate argument information, or semantic roles, to the syntactic structures of the Penn Treebank (Palmer et al., 2005). The Frame Files that provide guidance to the annotators constitute a rich English lexicon with explicit ties between syntactic realizations and coarse-grained senses, Framesets. PropBank Framesets are distinguished primarily by syntactic criteria such as differences in subcategorization frames, and can be seen as the top-level of an hierarchy of sense distinctions. Groupings of fine-grained WordNet senses, such as those developed for Senseval2 (Palmer et al., to appear) provide an intermediate level, where groups are distinguished by either syntactic or semantic criteria. WordNet senses constitute the bottom level. The PropBank Frameset distinctions, which can be made consistently by humans and systems (over 90% accuracy for both), are surprisingly compatible with the groupings; 95% of the groups map directly onto a single PropBank frameset sense (Palmer et al., 2004).

The semantic annotation provided by PropBank is only a first approximation at capturing the full richness of semantic representation. Additional annotation of nominalizations and other noun pred-
icates has already begun at NYU. This paper describes the results of PropBank II, a project to provide richer semantic annotation to structures that have already been propbanked, specifically, eventuality ID’s, coreference, coarse-grained sense tags, and discourse connectives. Of special interest to the machine translation community is our finding, presented in this paper, that PropBank II annotation reconciles many of the surface differences of the two languages.

2 PropBank I

PropBank (Palmer et al., 2005) is an annotation of the Wall Street Journal portion of the Penn Treebank II (Marcus et al., 1994) with ‘predicate-argument’ structures, using sense tags for highly polysemous words and semantic role labels for each argument. An important goal is to provide consistent semantic role labels across different syntactic realizations of the same verb, as in the window in [ARG0 John] broke [ARG1 the window] and [ARG1 The window broke]. PropBank can provide frequency counts for (statistical) analysis or generation components in a machine translation system, but provides only a shallow semantic analysis in that the annotation is close to the syntactic structure and each verb is its own predicate.

In PropBank, semantic roles are defined on a verb-by-verb basis. An individual verb’s semantic arguments are simply numbered, beginning with 0. Polysemous verbs have several framesets, corresponding to a relatively coarse notion of word senses, with a separate set of numbered roles, a role-set, defined for each Frameset. For instance, leave has both a DEPART Frameset ([ARG0 John] left [ARG1 the room]) and a GIVE Frameset, ([ARG0 I] left [ARG1 my pearls] [ARG2 to my daughter-in-law] [ARGM-LOC in my will].) While most Framesets have three or four numbered roles, as many as six can appear, in particular for certain verbs of motion. Verbs can take any of a set of general, adjunct-like arguments (ARGMs), such as LOC (location), TMP (time), DIS (discourse connectives), PRP (purpose) or DIR (direction). Negations (NEG) and modals (MOD) are also marked.

There are several other annotation projects, FrameNet (Baker et al., 1998), Salsa (Ellsworth et al., 2004), and the Prague Tectogrammatics (Haji-cova and Kuceroa, 2002), that share similar goals. Berkeley’s FrameNet project, (Baker et al., 1998; Fillmore and Atkins, 1998; Johnson et al., 2002) is committed to producing rich semantic frames on which the annotation is based, but it is less concerned with annotating complete texts, concentrating instead on annotating a set of examples for each predicate (including verbs, nouns and adjectives), and attempting to describe the network of relations among the semantic frames. For instance, the buyer of a buy event and the seller of a sell event would both be Arg0’s (Agents) in PropBank, while in FrameNet one is the BUYER and the other is the SELLER. The Salsa project (Ellsworth et al., 2004) in Germany is producing a German lexicon based on the FrameNet semantic frames and annotating a large German newswire corpus. PropBank style annotation is being used for verbs which do not yet have FrameNet frames defined.

The PropBank annotation philosophy has been extended to the Penn Chinese Proposition Bank (Xue and Palmer, 2003). The Chinese PropBank annotation is performed on a smaller (250k words) and yet growing corpus annotated with syntactic structures (Xue et al., To appear). The same syntactic alternations that form the basis for the English PropBank annotation also exist in robust quantities in Chinese, even though it may not be the case that the same exact verbs (meaning verbs that are close translations of one another) have the exact same range of syntactic realization for Chinese and English. For example, in (1), “新年/New Year招待会/reception” plays the same role in (a) and (b), which is the event or activity held, even though it occurs in different syntactic positions. Assigning the same argument label, Arg1, to both instances, captures this regularity. It is worth noting that the predicate “举行/hold” does not have passive morphology in (1a), despite what its English translation suggests. Like the English PropBank, the adjunct-like elements receive more general labels like TMP or LOC, as also illustrated in (1). The functional tags for Chinese and English PropBanks are to a large extent similar and more details can be found in (Xue and Palmer, 2003).

(1) a. [ARG1 新年/New Year 招待会/reception] [ARGM-TMP 今天/today] [ARGM-LOC 在/at 钓鱼

62
The New Year reception was held in Diaoyutai State Guest House today.

Tang Jiaxuan was holding the New Year reception in Diaoyutai State Guest House today.

(2) a. Mr. Bush met him privately in the White House on Thursday.

b. Propbank I: Rel: met, Arg0: Mr. Bush, Arg1: him, ArgM-MNR: privately, ArgM-LOC: in the White House, ArgM-TMP: on Thursday.

c. Propbank II: ∃e meeting(e) & Arg0(e, Mr. Bush) & Arg1(e, him) & MNR(e, privately) & LOC(e, in the White House) & TMP(e, on Thursday).

Annotation of event variables starts by automatically associating all PropBank I annotations with potential event ids. Since not all annotations actually denote eventualities, we manually filter out selected classes of verbs. We further attempt to identify all nouns and nominals which describe eventualities as well as all sentential arguments of the verbs which refer to events. And, finally, part of the PropBank II annotation involves tagging of event coreference for pronouns as well as empty categories. All these tasks are discussed in more detail below.

Identifying event modifiers. The actual annotation starts from the presumption that all verbs are events or states and nouns are not. All the verbs in the corpus are automatically assigned a unique event identifier and the manual part of the task becomes (i) identification of verbs or verb senses that do not denote eventualities, (ii) identification of nouns that do denote events. For example, in (3), begin is an aspectual verb that does not introduce an event variable, but rather modifies the verb ‘take’, as is supported by the fact that it is translated as an adverb ”initially” in the corresponding Chinese sentence.

Nominalizations as events Although most nouns do not introduce eventualities, some do and these nouns are generally nominalizations. This is true...
for both English and Chinese, as is illustrated in (4). Both “发展/develop” and “深入/deepening” are normalized verbs that denote events. Having a parallel propbank annotated with event variables allows us to see how events are lined up in the two languages and how their lexical realizations can vary. The normalized verbs in Chinese can be translated into verbs or their nominalizations, as is shown in the alternative translations of the Chinese original in (4). What makes this particular example even more interesting is the fact that the adjective modifier of the events, “不断/continued”, can actually be realized as an aspectual verb in English. The semantic representations of the Propbank II annotation, however, are preserved: both the aspectual verb “continue” in English and the adjective “不断/continued” in Chinese are modifiers of the events denoted by “发展/development” and “深 入/deepening”.

(4) 随着/with 中国/China 经济/economy 的/DE 不断/continued 发展/development 和/to 外/outside 开放/open 的/DE 不断/continued 深入/deepen ...
“As China’s economy continues to develop and its practice of opening to the outside continues to deepen…”

“With the continued development of China’s economy and the continued deepening of its practice of opening to the outside…”

Event Coreference Another aspect of the event variable annotation involves identifying pronominal expressions that corefer with events. These pronominal expressions may be overt, as in the Chinese example in (5), or non-overt. In (6), the null pronoun in Chinese is translated into an overt pronoun in English.

(5) 去年/last year, 纽约/New York 所/new 上市/list 的/DE 外国/foreign 企业/enterprise 等/altogether 有/with 61/61 家/CL, *pro* 创建/create 历年/recent year 来/since 最高/highest 纪录/record.
“Last year, there were 61 new foreign enterprises listed in New York Stock Exchange, *PRO* creating the highest record in history.”

Having a parallel propbank annotated with event variables allows us to examine how the same events are lexicalized in English and Chi-nese and how they align, whether they have been indicated by verbs or pronouns.

3.2 Grouped sense tags
In general, the verbs in the Chinese PropBank are less polysemous than the English PropBank verbs, with the vast majority of the lemmas having just one Frameset. On the other hand, the Chinese PropBank has more lemmas (including static verbs which are generally translated into adjectives in English) normalized by the corpus size. The Chinese PropBank has 4854 lemmas in the 250K words that have been propanked alone, while the English PropBank has just 3635 lemmas in the entire 1 million words corpus. Of the 4854 Chinese lemmas, only 62 of them have 3 or more framesets. In contrast, 294 lemmas have 3 or more framesets in the English PropBank.
| Verb | English senses | Chinese translations |
|------|----------------|----------------------|
| appear | be or have a quality of being | 显得,呈现 |
|       | come forth, become known or visible, physically or figuratively | 出现,呈现 |
|       | present oneself formally, usually in a legal setting | 露面 |
| fight | combat or oppose | 打好, 打, 打抗 |
|       | strive, make a strenuous effort | 奋斗 |
|       | promote, campaign or crusade | 奋斗 |
| join | connect, link or unite separate things, physically or abstractly | 链接, 接轨 |
|       | enlist or accept membership within some group or organization | 走进, 参加, 加入 |
|       | participate with someone else in some event | 同...一道, 同...一起 |
| realize | be cognizant of, comprehend, perceive | 认识, 意识 |
|       | actualize, make real | 实现 |
|       | take in, earn, acquire | 实入 |
| pass | travel by | 经 |
|       | clear, come through, succeed | 过去, 成功 |
|       | elapse, happen | 过去, 发生 |
| communicate | | 传出 |
| settle | resolve, finalize, accept | 决定, 完成 |
|       | reside, inhabit | 进驻, 落户 |
| raise | increase | 提高 |
|       | lift, elevate, orient upwards | 仰 |
|       | collect, levy | 收集, 征集, 征措 |
|       | invoke, elicit, set off | 起, 提出 |

Table 1: English verbs and their translations in the parallel Propbank

In our sense-tagging part of the project, we have been using manual groupings of the English WordNet senses. These groupings were previously shown to reconcile a substantial portion of the tagging disagreements, raising inter-annotator agreement from 71% in the case of fine-grained WordNet senses to 82% in the case of grouped senses for the Senseval 2 English data (Palmer et al., to appear), and currently to 89% for 93 new verbs (almost 12K instances) (Palmer et al., 2004). The question which arises, however, is how useful these grouped senses are and whether the level of granularity which they provide is sufficient for such applications as machine translation from English to Chinese.

In a preliminary investigation, we randomly selected 7 verbs and 5 nouns and looked at their corresponding translations in the Chinese Propbank. As the tables below show, for 6 verbs (join, pass, settle, raise, appear, fight) and 3 nouns (resolution, organization, development), grouped English senses map to unique Chinese translation sets. For a few examples, which include realize and party, grouped senses map to the same word in Chinese, preserving the ambiguity. This investigation justifies the appropriateness of the grouped sense tags, and indicates potential for providing a useful level of granularity for MT.

3.3 Discourse connectives

Another component of the Chinese / English Parallel Propbank II is the annotation of discourse connectives for both Chinese corpus and its English translation. Like the other two components, the annotation is performed on the first 100K words of the Parallel Chinese English Treebank. The annotation of Chinese discourse connectives follows in large part the theoretic assumptions and annotation practices of the English Penn Discourse Project (PDTB) (Miltsakaki et al., 2004). Adaptations are made only when they are warranted by the linguistic facts of Chinese. While the English PTDB annotates both explicit and implicit discourse connectives, our ini-
tial focus is on explicit discourse connectives. Explicit discourse connectives include subordinate (8) and coordinate conjunctions (9) as well as discourse adverbials (10). While subordinate and coordinate conjunctions are easy to understand, discourse adverbials need a little more elaboration. Discourse adverbials differ from other adverbials in that they relate two propositions. Typically one can be found in the immediate context while the other may need to be identified in the previous discourse. (8) [arg1 台湾/Taiwan 商人/businessman] [conn 虽然/although] [arg2 他们/their live 生活/live 在/at 外/foreign land], [arg2 还是/still 保/保持/very 重视/stress 孩/孩子 教育/education]。“Although these Taiwan businessmen live away from home, they still stress the importance of their children’s education.”

(9) [arg1 东亚/East 合/合作/and 分歧/difference], [conn 但是/but] [arg2 为了/for 保障/protect 东亚/East Asia 合/合作/and 分歧/difference], [arg2 为了/for 保障/protect 东亚/East Asia 利/利益/interest, 必须/must 进一步/progress 更进/forward 加强/strengthen 合作/合作/and 分歧/difference]。“It is not really true that there are no conflicts and differences among the East Asian countries, but in order to protect their common interest, they must cooperate.”

(10) [arg1 浦东/Pudong 开发/development 是/IS BE 一/one 项/ONE 项目/PROJECT, [conn 为了/for 提升/rapidly/improve 跨/over/世纪/century 工程/PROJECT], [arg2 大量/large quantity 问题/problem, “The development of Pudong, a project designed to invigorate Shanghai, spans over different centuries. Therefore, new problems occur in large quantities.”

Table 2: English nouns and their translations in the parallel Propbank

| Noun   | English senses | Chinese translations |
|--------|---------------|---------------------|
| organization | individuals working together | 组织,机构,单位 |
| event: putting things together | 状态:被组织性 |
| party | event: an occasion on which people can assemble for social interaction and entertainment | 会 |
| political organization | 政治组织 |
| a band of people associated temporarily in some activity | 一方 |
| person or side in legal context | 法律 |
| investment | time or money risked in hopes of profit | 投资,资金 |
| development | the process of development | 开发,发展 |
| resolution | a formal declaration | 协议,决定 |
| coming to a solution | 解决 |

The annotation of the discourse connectives in a parallel English Chinese Propbank exposes interesting correspondences between English and Chinese discourse connectives. The examples in (11) show that “结果” is polysemous and corresponds with different expressions in English. It is a noun meaning “result” in (11a), where it is not a discourse connective. In (11b) it means “in the end”, invoking a contrast between what has been planned and how the actual result turned out. In (11c) it means “as a result”, expressing causality between the cause and the result.

(11) a. 实行/adopt “我民愿用老/go slow” 的/DE 政策/policy, 结果/result 是/BE 等于/unnecessarily 丢失/lose 在/at 大陆/mainland 的/DE 机会/business opportunity.”

“The result of adopting the ‘go slow’ policy is unnecessarily losing business opportunities in the mainland.”

b. 纤维所/fiber institute 计划/plan 招收/enroll 十/10 名/COUNT 学生/student, 结果/result 在/in the end 只/only 有/have 三/3 人/person 报名/register.”

“The fiber institute planned to enroll 10 students. In the end, only 20 people registered to take the exam.”

c. 学校/school 不/not 教教理财政管理, 一般/ordinary 人/people 为/for 有/have 我/this 方面/aspect 的/DE 需求/need, 结果/result as a result, 报章/newspaper 上/on 各/every 种/kIND 专栏/column 然/then 成为/become 资讯/information 的/DE 主要/main 来源/source.”

“The school does not teach finance management and
ordinary people have this need. As a result, the different kinds of columns in the newspaper become the main source of information.”

4 Conclusion

This paper presented preliminary results of the parallel PropBank II project. It highlighted some interesting aspects of the differences between English and Chinese, which play an important role for MT and other applications. Some of the questions addressed had to do with how events are lexicalized and aligned in the two languages, which level of sense granularity is needed for MT from English to Chinese, and highlighting notable differences between discourse connectives in the two languages. Further investigation and alignment of the parallel corpus, as well as richer annotation, will reveal other interesting phenomena.

References

Olga Babko-Malaya and Martha Palmer. 2005. Proposition Bank II: Delving Deeper. In *Frontiers in Corpus Annotation, Workshop in conjunction with HLT/NAACL 2004*, Boston, Massachusetts.

Olga Babko-Malaya, Martha Palmer, Nianwen Xue, Aravind Joshi, and Seth Kulick. 2004. Exploiting Interactions between Different Types of Semantic Annotation. In *Proceedings of ICWS-6*, Tilburg, The Netherlands.

C. Baker, C. Fillmore, and J. Lowe. 1998. The berkeley framenet project. In *Proceedings of COLING-ACL*, Singapore.

E. Charniak. 2001. Immediate-head Parsing for Language Models. In *ACL-01*.

Michael Collins. 1999. *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

M. Ellsworth, K. Erk, P. Kingsbury, and S. Pado. 2004. PropBank, SALS A and FrameNet: How design determines product. In *Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Lisbon, Portugal.

Charles J. Fillmore and B. T. Atkins. 1998. FrameNet and lexical relevance. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.

Eva Hajicova and Iyona Kucerova. 2002. Argument/Valency Structure in PropBank, LCS Database and Prague Dependency Treebank: A Comparative Pilot Study. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 846–851.

Christopher R. Johnson, Charles J. Fillmore, Miriam R. L. Petruck, Collin Baker, Michael Ellsworth, Josef Ruppenhofer, and Esther J. Wood. 2002. FrameNet: Theory and Practice, Version 1.0, www.icsi.berkeley.edu/fframenet.

M. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, et al. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In *Proc of ARPA speech and Natural language workshop*.

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Martha Palmer, Olga Babko-Malaya, and Hoa Dang. 2004. Different Sense Granularities for Different Applications. In *Proceedings of the 2nd Workshop on Scalable Natural Language Understanding Systems*, Boston, Mass.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. to appear. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering*.

Nianwen Xue and Martha Palmer. 2003. Annotating the Propositions in the Penn Chinese Treebank. In *The Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

Nianwen Xue, Fei Xia, Fu dong Chiou, and Martha Palmer. To appear. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*.

Nianwen Xue. To appear. Annotating the Discourse Connectives in the Chinese Treebank. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation*, Ann Arbor, Michigan.