

Automatic Predicate Argument Structure Analysis of the Penn Chinese Treebank

Nianwen Xue and Seth Kulick

{xueniwen, skulick}@linc.cis.upenn.edu
Dept. of Computer and Information Science
University of Pennsylvania
200 S. 33rd Street
Philadelphia, PA 19104, USA

Abstract

Recent work in machine translation and information extraction has demonstrated the utility of a level that represents the predicate-argument structure. It would be especially useful for machine translation to have two such Proposition Banks, one for each language under consideration. A Proposition Bank for English has been developed over the last few years, and we describe here our development of a tool for facilitating the development of a Chinese Proposition Bank. We also discuss some issues specific to the Chinese Treebank that complicate the matter of mapping syntactic representation to a predicate-argument level, and report on some preliminary evaluation of the accuracy of the semantic tagging tool.

1 Introduction

Recent work in machine translation has focused on the level of predicate-argument relations as an appropriate interlingual representation for machine translation (Hajic et al., 2002). This work however has been limited by the lack of corpora in the desired languages that are annotated with these relations.

The creation of structurally annotated treebanks (Marcus et al., 1993; Marcus et al., 1994) led to major advances in corpus-based natural language technologies. Most notably, the syntactically annotated corpora have proven to be a crucial resource

in the recent success of the statistical natural language parsers (Collins, 1997; Collins, 2000; Charniak, 2000), and the development of the Penn Chinese Treebank (Xia et al., 2000) is also beginning to help advance technologies in Chinese syntactic analysis (Chiang, 2000; Bikel and Chiang, 2000).

However, since these treebanks are syntactically oriented, significant regularities in the predicate-argument structure of lexical items are not captured. The development of the Penn PropBank (Kingsbury and Palmer, 2002) is beginning to address this issue for English. In this new layer of annotation, the regularities of the predicates (currently mostly verbs), are captured in an explicit level of predicate-argument annotation. This annotation has led to the development of semantic taggers (Gildea and Palmer, 2002). Recent work in information extraction (Surdeanu et al., 2003) has shown how this can be used to map tagged semantic roles to template slots, eliminating the need for complex pattern-matching rules.

This approach can be extended to machine translation tasks only if there exists a corpus for the second language that is similarly annotated for predicate-argument structure. For example, for English-Chinese translation, Propbanks for both English and Chinese corpora would allow the training of semantic taggers for both languages. The MT work would then involve the mapping of the corresponding predicate-argument structures, much as the IE work using the Propbank mapped the predicate-argument structure to a template.

In this work we report on our development of a tool for facilitating the development of a Chinese

Propbank. While a good deal of the work has to be done by human annotators, the tagging software we discuss here attempts a first-pass at the semantic annotation, to be then corrected by the human annotators. The software uses the structural information in the Chinese Treebank to determine the Propbank annotation. It is of course impossible for this job to be done 100% automatically (if it were possible, then there would be no need for the Propbank). However, a certain amount of information can be automatically determined. There are also certain aspects of the Chinese parses, discussed below, that are different from those for English and that complicate the mapping to the predicate argument structure.

The structure of the paper is as follows. Section 2 discusses in more detail the motivation for a Chinese Propbank, and its relevance for Machine Translation. Section 3 discusses the annotation scheme used for the Propbank, and Section 4 discusses how the automatic semantic tagger can use the syntactic information in the Treebank to automatically create the Propbank. Section 5 discusses some preliminary results of an evaluation of the success of the automatic tagging, and section 6 is the conclusion.

2 Motivation

In this section we demonstrate the utility of multilingual Propbanks, focusing on their relevance to Machine Translation. In the framework set up in the Penn Propbank (Kingsbury and Palmer, 2002), a predicate is split into multiple senses (called *rolesets*) if it has different predicate argument structures. For example, the verb *sign* has two *rolesets*: *sign.01* and *sign.02*¹. *sign.01*, meaning “affix a signature”, takes three arguments: *signer*, *document* and *signature*. *sign.02*, meaning “enter an agreement”, takes three different arguments, *agreeer*, *agreement* and *co-signer*. As is generally expected, there is no single translation for the verb *sign* in Chinese. At least four verbs are possible translations and they are 签订, 签字, 签署 and 签. However, not all four Chinese verbs are possible translations for both senses of the English verb *sign*. For example, while 签 can be a translation of both senses of *sign*, 签订 can only be a translation of *sign.02* and 签字, 签署 can only be used to translate *sign.01*. Splitting a verb into multiple *rolesets* based

on its predicate argument structure opens up the possibility of linking a *roleset* of a predicate in one language with that of another, rather than using bilingual (or multilingual) dictionaries where one word is translated into one or more words in a different language. Having a bilingual lexical resource where verbs are listed with its *rolesets* and their predicate argument structures has several advantages, compared with an ordinary bilingual dictionary. First of all, mapping at the *roleset* level is generally more precise than mapping at the word level. For example, the possible translations of *sign.01* are 签 (1a), 签字 (1b) and 签署 (1c) (to the exclusion of 签订) while the possible translations of *sign.02* are 签 (2a) and 签订 (2b) (to the exclusion of 签字 and 签署). Assuming English is the source language and Chinese is the target language, this means that in an actual occurrence of English verb *sign*, the identification of its *roleset* will narrow down the possible lexical choices in Chinese.

For a given instance of the English verb *sign*, it is not always the case that all of its possible arguments are realized. For example, both “he signed the document” and “he signed his name on the document” are legitimate English sentences even though in the former sentence the *signature* argument is missing. Depending on which arguments are realized in the source language, the lexical choices in the target language may be different, because not all possible translations of a *roleset* take the same arguments. For example, if the realized arguments of *sign.01* are the signer, the document, and the signature, as in (1a), among the translation candidates 签署, 签字 and 签, only 签 is possible, because both 签署 (1c) and 签字 (1b) can only take two arguments, namely, the signer and the document. Recognizing the arguments of a predicate instance is thus crucial in further constraining the lexical choices and rule out translation candidates whose predicate argument structures are incompatible.

- 1a. 他/he 在/at 这/this 个/CL 文件/document 上/LC
签/sign 了/LE 自己/self 的/DE 名字/name
“He signed his name on this document.”
- 1b. 他/he 在/at 这/this 个/CL 文件/document 上/LC
签字/sign
“He signed this document.”
- 1c. 他/he 签署/sign 这/this 个/CL 文件/document

¹See www.cis.upenn.edu/~ace.

“He signed this document.”

- 2a. 中国/China 和 美国/the U. S. 签/sign 了/LE
协议/agreement.
“China and U.S.A. signed an agreement.”
- 2b. 中国/China 和 美国/U.S.A. 签订/sign 协议/agreement.
“China and U.S.A. signed an agreement.”

One might argue that the syntactic subcategorization frame obtained from the syntactic parse tree in a treebank can also constrain the lexical choices. For example, knowing that *sign* has a subject, an object and a prepositional phrase (e.g. 1a) should be enough to rule out 签署 and 签字 as possible translations, because the former can only take a subject and an object (1c) while the latter can only take a subject and a prepositional phrase (1b). This argument breaks down when there are lexical divergences. In the English sentence “he signed this document”, the “document” is realized as the direct object. If the syntactic subcategorization frame is used to constrain the lexical choices for “sign”, “签字” will be incorrectly ruled out as a possible translation because the “document” argument of “签字” can only be realized as a prepositional phrase. There would be no such problem if the more abstract predicate-argument structure is used for this purpose. Even when the document is realized as a prepositional phrase, it is still the same argument. Of course, “签署/sign” is also a possible translation. So compared with the surface syntactic frames, the predicate-argument structure constrains the lexical choices without incorrectly ruling out legitimate translation candidates. This is understandable because the predicate-structure abstracts away from the syntactic idiosyncrasies of the different languages and thus are more transferable across languages.

In order to recognize the predicate argument structure it is crucial to create a Propbank in which the predicates are tagged with *rolesets* and argument structures so that the automatic semantic analyzers can be trained and tested. A sensible approach will be to bootstrap from an existing treebank, where sentences are annotated with syntactic structures. In the next section, we will briefly describe the Penn Chinese Treebank and demonstrate how we can use the syntactic information to automatically tag the

predicate argument structure.

3 From Syntactic Annotation to Semantic Annotation

The Penn Chinese Treebank (Xia et al., 2000) (CTB) is a segmented, POS-tagged and syntactically bracketed corpus consisting of articles from a variety of sources: Xinhua newswire, the Hong Kong News, and Sinorama. The syntactic entities for each sentence are marked with a combination of hierarchically organized labeled brackets, functional tags, null elements and indices. An example is given in (1).² It is clear that with this type of annotation, the subject and object for the verb 通过 (“pass”) can be easily extracted.

3. (IP (NP-PN-SBJ 美国/the U.S. 国会/Congress)
(VP (PP-TMP 于/at 昨日/yesterday)
(VP (VV 通过/pass)
(AS 了/ASP)
(NP-OBJ 州际/interstate
銀行法/banking law))))

The U.S. Congress passed the interstate banking law.

In annotating the arguments for the predicates in the CTB we follow the conventions set up in (Kingsbury and Palmer, 2002). The arguments for a given predicate are assigned semantic labels in the form of *argN*, where *N* is an integer between 0 and 5. For instance, for the verb 通过 (“pass”) in (3), the subject 美国 国会 (“the U.S. Congress”) will receive the label *arg0* while the object 州际 銀行法 (“the interstate banking law”) will be given the label *arg1*. The adjuncts will be represented as *argM*, with a functional tag indicating the type of adjunct. For instance, 于 昨日 (“yesterday”) will be annotated as *argM-TMP*. This type of annotation assumes that each predicate is associated with a limited set of semantic roles that are assigned to its arguments. Each semantic role is unique with respect to the predicate, although there is no attempt to generalize the semantic role labels across different predicates³. On one hand, by not

²The glosses and translations are not part of the CTB annotation and the structures in the examples are simplified.

³In practice, it is possible that the same semantic role label has a similar denotation across predicates of a particular class, or even across all the predicates, but that is not necessarily the case.

using notions such as “agent” or “theme” that often carry a global meaning generalizable across all predicates, this representation scheme avoids the thorny issue of deciding whether arguments with the same semantic label actually share certain semantic properties. On the other hand, this representation scheme captures the regularities across different instances of a given predicate that the treebank annotation does not capture. For instance, in the example in (4), even though **州际 银行法** (“the interstate banking law”) is represented as the subject of **通过** (“pass”) instead of the object, in this sentence, it is intuitively clear that it plays the same semantic role as the object in (3). Therefore, instead of receiving the semantic role label *arg0*, it should be assigned the label *arg1*.

4. (IP (NP-SBJ (ADJP (JJ 州际/interstate)) (NP (NN 银行法/banking law))) (VP (VV 通过/pass) (AS 了/ASP)))
The interstate banking law passed.

Several observations can be made about the relation between the syntactically-oriented annotation in the CTB and the proposed semantically-oriented propbank annotation. First of all, there is no one-to-one mapping between the syntactic entities such as the subject and the semantic entities such as *arg0*. A given semantic entity can be realized in different syntactic positions. For example, the *arg1* of **通过** (“pass”) can occur in both the object and the subject positions. Second, not all semantic roles associated with a predicate need to be realized syntactically for a given instance of a predicate. For example, the sentence in (4) does not specify who passed the interstate banking law, although it is still understood that the law must be passed by some political entity. In other words, the “passer” is still understood as assuming the role *arg0* even if it is not explicitly expressed. All the possible syntactic realizations of a certain predicate can be said to be alternations of one another. Levin (1993) has argued convincingly that such alternations are systematic and signal underlying semantic similarities between English verbs.

In addition to such diathesis alternations, arguments bearing the same semantic role may also end up in different syntactic positions as a result of dislocation or movement of a more global nature (e.g.,

wh-movement). Instances of dislocation are clearly represented in the CTB with traces and indices. Using the structural information in the CTB as well as the alternation patterns for classes of verbs, we are able to map the syntactic structure to the predicate argument structure, as we will demonstrate in the next section.

4 A Semantic Tagger Based on Parse Trees

We describe in this section the basis for a semantic tagger, which provides a first approximation of the propbank semantic annotation, to be later corrected by human annotators. This tagger therefore plays a role in the creation of the Chinese Propbank similar to that of the system described in (Palmer et al., 2001) for the English Propbank. A major difference however is that the tool for the English Propbank takes advantage of lexical resources such as the Susanne corpus and WordNet, while such resources are not available for Chinese. Since it is also not a simple matter to induce lexical diathesis information from the Treebank, we manually provide some such information for the semantic tagger.

To carry out this first approximation to mapping the arguments to semantic roles, we extract out of the CTB the structural information for each predicate. In most cases this is straightforward, consisting of traversing the parse tree from each verb to the top of that verb’s clause, picking up all the complements and subject of the verb.

For example, the sentence in (3) would produce the output:

NP-PN-SBJ: 美国/the U.S. 国会/Congress PP-TMP: 于/at 昨日/yesterday REL: 通过/pass NP-OBJ: 州际/interstate 银行法/banking law
--

Using this sort of simple structural approach to recover the predicate-argument structure allowed us to capture the simple syntactic cases, namely all the ones in which there is sufficient syntactic annotation. This includes cases of syntactic movement, such as wh-movement or topicalization.

However, two constructions that are quite frequent in the CTB are not so straightforward. These are particularly interesting since in one case a syn-

tactic bi-clausal structure is made into a monoclausal subcategorization frame, while in the other a syntactic monoclausal structure may result in two subcategorization frames.

4.1 BA and BEI Constructions

An example of the first is the *ba* construction in Chinese, with the pattern NP0 + BA + NP1 + VP, which means “NP0 causes NP1 to maintain a state or to change to a new state”.

The way the treebank was constructed, *ba* is a verb taking an IP clause as a complement.⁴ However, although the *ba* constructions consist of two clauses, with two verbs (including *ba*), each such construction results in one subcategorization frame. The following Example (5) is an example of this from the treebank:

5. (IP (NP-SBJ 卫生部/health Ministry)
 (VP (BA 将)
 (IP-OBJ (NP-SBJ 其/this)
 (VP 定为/designate
 (NP-OBJ 国家级/state-level
 新/new
 药/medicine))))))

The Health Ministry has designated this as state-level new medicine.

Subcat Frame:

NP-SBJ: 卫生部/Health Ministry
 NP-OBJ: 其/this
 REL: 定为/designate
 NP-OBJ2: 国家级/state-level 新/new 药/medicine

There is also a version of the *ba* construction in which the lower verb has no object. In this case the subject of the lower clause is a displaced object of the lower verb, although such trace information is not explicitly represented in the CTB. This is therefore somewhat like a passive in English, although there is no passive morphology. Example (6) is an illustration of this, and it again results in a single subcategorization frame.

6. (IP (NP-PN-SBJ 大荣/Darong
 公司/company)
 (VP (BA 把/ba)
 (IP-OBJ (NP-SBJ 中国/Chinese
 商品/commodity
 采购量/purchase)
 (VP (VV 扩大/increase))))))

Darong Company increased its purchase of Chinese commodities.

Subcat Frame:

NP-PN-SBJ: 大荣/Darong 公司/company
 NP-OBJ: 中国/Chinese 商品/commodity
 采购量/purchase volume
 REL: 扩大/increase

4.2 Verb Compounds

The other major case of a syntax/semantics mismatch is that of verb compounds in the Chinese Treebank. In the CTB, verb compounds can be loosely defined as several verbs sharing the same arguments. For example, the VCD compound indicates that the verb constituents of the compound share the same arguments, and so have the same subcategorization frame. This is illustrated in (7).

7. (IP (NP-SBJ 我们/we)
 (VP (VV 要/should)
 (VP (VCD (VV 贯彻/follow)
 (VV 落实/implement)))
 (NP-OBJ (NP (NP (DP 全/whole)
 (NP 国/nation))
 (NP 经济/economic
 工作/work
 会议/meeting))
 (NP 精神/spirit))))))

We should follow and implement the spirit of the national economic work meeting

The sentence actually produces two subcategorization frames, one for each of the verbs in the VCD compounds:

⁴The *bei* construction raises issues similar to that of *ba*, although here we discuss only *ba*.

Subcat Frame 1:

NP-SBJ: 我们/we

REL: 贯彻/follow

NP-OBJ: 全/whole 国/nation 经济/economic
工作/work 精神/spirit**Subcat Frame 2:**

NP-SBJ: 我们/we

REL: 落实/implement

NP-OBJ: 全/whole 国/nation 经济/economic
工作/work 精神/spirit

However, not all verb compounds result in different subcat frames. For example, VSB is the label for a verb compound forming a modifier/head relationship. In this case we only get a subcat frame for the right constituent (the head). An example from the treebank is given in (8):

8. (IP (NP-SBJ (-NONE- *PRO*))
 (VP (VSB (VV 前来/come forward)
 (VV 投资))))

(Foreign investors) come to invest

Subcat Frame:

NP-SBJ: *PRO*

REL: 投资/invest

There are a variety of other verb compounds, which we cannot run through in detail here, but we have determined for each of the different compounds, whether and how to break up the constituents of the compounds into different subcat frames.

Having extracted the subcat frames of the verbs, we are ready to map them to predicate argument structures.

5 Experiment and Discussion

In this section we describe experimental results of assigning semantic labels of the type described in Section 3, using the structural information in the CTB as well as diathesis alternation information of the verbs. Ideally we would need a list of Chinese verbs with their alternation patterns specified. Since to our knowledge there is no readily available source of such information, we manually selected 30 verbs that demonstrate “object of transitive / subject of intransitive” alternation and extracted all instances of them from the CTB (250K words worth of data),

as described in Section 4. The subcategorization frames for each instance of the verbs are extracted and their arguments are automatically assigned semantic labels, using the alternation pattern for this type of verbs. That is, the subject is *arg0* when the verb is transitive and *arg1* when the verb is intransitive while the object is always assigned *arg1*. For purposes of this experiment, we did not attempt to determine the *rolesets* and tag the adjuncts. The accuracy of the mapping from syntactic entities to semantic roles is calculated on a “per-predicate” basis. That is, if any of the arguments of a verb is assigned the wrong label, that instance of the verb is considered an error. The accuracy for a predicate is simply the number of correctly tagged instances divided by the total number of instances of this predicate. The accuracy is calculated in two ways. One score takes into consideration all instances of the word when used as a predicate and another score only considers instances of the word used as a verb. In order to determine the nature of the errors we break down the errors for transitive and intransitive uses of the verbs. The results of our experiments are tabulated in Table 1.

The results show that with the structural information provided in the CTB and prior knowledge of the diathesis alternation information 95% of the verb instances are tagged correctly. An error analysis makes clear three major issues that require further study: (i) Some verbs cannot be simply characterized as having “subject of intransitive/object of transitive” alternation because different senses of the same verb may have different alternation patterns. (ii) While the results based on just the verb instances are quite satisfactory, if nominalizations are also taken into account, the accuracy drops to 79%. We did not attempt to annotate the predicate-argument structure for nominalized verbs because the annotation in the CTB does not allow straightforward mapping from syntactic representation to semantic representation. (iii) Another type of error is due to what can be characterized as syntax/semantics mismatches. A common source of error can be attributed to the fact that some semantic arguments are realized as syntactic adjuncts, so that reading the arguments off a syntactic parse tree is no longer straightforward.

The results obtained here are considerably higher

word	gloss	pred	noun	verb	accuracy (verb/all)	err	intr. err.	tr. err.
缓解	alleviate	13	1	12	1/0.92	0	0	0
改变	change	30	5	25	0.84/0.7	4	4	0
建成	build-into	129	16	113	0.89/0.78	12	0	12
宣布	declare	87	4	83	0.95/0.91	4	0	4
加深	deepen	8	2	6	1/0.75	0	0	0
深化	deepen	23	0	23	1/1	0	0	0
分散	disperse	4	0	4	1/1	0	0	0
缓和	ease	9	4	5	1/0.56	0	0	0
扩大	expand	160	7	153	0.90/0.86	15	1	3
建立	establish	175	7	168	0.93/0.89	12	0	0
成立	establish	134	29	105	0.99/0.78	1	0	0
普及	popularize	8	1	7	0.86/0.75	1	0	1
拓宽	widen	7	0	7	1/1	0	0	0
举行	hold	229	2	227	0.99/0.98	1	0	1
改善	improve	79	27	52	0.98/0.65	1	0	1
加剧	intensify	3	0	3	1/1	0	0	0
开通	open	33	6	27	1/0.82	0	0	0
优化	optimize	27	6	21	0.95/0.74	1	0	1
通过	pass	78	5	73	0.93/0.87	5	4	0
完善	perfect	53	5	48	1/0.91	0	0	0
产生	generate	33	3	30	0.87/0.79	4	3	1
繁荣	flourish	47	34	13	0.92/0.26	1	1	0
解除	terminate	19	0	19	1/1	0	0	0
化解	resolve	9	2	7	1/0.78	0	0	0
牺牲	sacrifice	6	4	2	1/0.33	0	0	0
签订	sign	53	1	52	0.58/0.57	22	0	22
巩固	solidify	16	2	14	0.93/0.81	1	0	1
解决	solve	124	23	101	0.99/0.81	1	0	0
启动	start	51	17	34	0.94/0.63	2	1	1
动摇	waiver	6	0	6	1/1	0	0/0	0/0
Average		0.95/0.79						

Table 1: Experimental results

than the 83.7% and 82.8% reported in (Palmer et al., 2001) and (Gildea and Palmer, 2002) respectively, where the data is from the Penn English Treebank and the results are reported on verbs only. However, the results are not directly comparable for several reasons. First, Chinese verbs seem to have fewer senses and thus fewer alternation patterns than English. As a result, the mapping between predicate argument structures is more likely to be deterministic. Second, the CTB distinguishes syntactic complements and adjuncts structurally (Xue and Xia, 2000) while the Penn English Treebank does not and this information turns out to be very useful for extracting the subcategorization frames that are mapped onto the predicate argument structures.

6 Conclusion and Future Work

We have argued here for the need for a level of predicate-argument annotation on top of the current Chinese Treebank. Machine translation by mapping of the predicate-argument structures helps to overcome the ambiguity that plagues MT efforts based at a word level. We presented the basic annotation scheme for the Chinese Propbank, and discussed how a first approximation to this annotation can be automatically derived from the Chinese Treebank. There are several major issues involved in such a derivation, as is to be expected since the Treebank is not sufficient for recovery of the predicate-argument structure. However, as our preliminary evaluation of this work has shown, it should be very useful in speeding up production of the Chinese Propbank. At the same time, the very process of uncovering such problems helps pinpoint the areas where the Propbank will be particularly essential for information extraction and machine translation purpose. In future work we will address the Propbank annotation of these problematic cases, as well as extend the automatic tagging to a wider range of semantic roles.

7 Acknowledgement

This work is supported by MDA904-02-C-0412.

References

- Eugene Charniak. 2000. A Maximum-Entropy-Inspired Parser. In *Proc. of NAACL-2000*.
- Mike Collins. 1997. Three Generative, Lexicalised Models for Statistical Parsing. In *Proc. of ACL-1997*.
- Mike Collins. 2000. Discriminative Reranking for Natural Language Parsing. In *Proc. of ICML-2000*.
- Daniel Gildea and Martha Palmer. 2002. The Necessity of Syntactic Parsing for Predicate Argument Recognition. In *Proc. of ACL-2002*.
- Jan Hajič, Martin Čmejrek, Bonnie Dorr, Yuan Ding, Jason Eisner, Daniel Gildea, Terry Koo, Kristen Parton, Gerald Penn, Dragomir Radev, and Owen Rambow. 2002. Natural language generation in the context of machine translation. Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore. Summer Workshop Final Report.
- Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, Las Palmas, Spain.
- Beth Levin. 1993. *English Verbs and Alternations: A Preliminary Investigation*. Chicago: The University of Chicago Press.
- 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*.
- Martha Palmer, Joseph Rosenzweig, and Scott Cotton. 2001. Automatic predicate argument analysis of the penn treebank. In *Proceedings of HLT 2001, First International Conference on Human Language Technology Research*, San Francisco.
- Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL 2003*, Sapporo, Japan.
- Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. In *Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece.
- Nianwen Xue and Fei Xia. 2000. The Bracketing Guidelines for Penn Chinese Treebank Project. Technical Report IRCS 00-08, University of Pennsylvania.