Treebank of Chinese Bible Translations

Andi Wu
GrapeCity Inc.
andi.wu@grapecity.com

Abstract

This paper reports on a treebanking project where eight different modern Chinese translations of the Bible are syntactically analyzed. The trees are created through dynamic treebanking which uses a parser to produce the trees. The trees have been going through manual checking, but corrections are made not by editing the tree files but by re-generating the trees with an updated grammar and dictionary. The accuracy of the treebank is high due to the fact that the grammar and dictionary are optimized for this specific domain. The tree structures essentially follow the guidelines of the Penn Chinese Treebank. The total number of characters covered by the treebank is 7,872,420 characters. The data has been used in Bible translation and Bible search. It should also prove useful in the computational study of the Chinese language in general.

1 Introduction

Since the publication of the Chinese Union Version (CUV 和合本) in 1919, the Bible has been re-translated into Chinese again and again in the last 91 years. The translations were done in different time periods and thus reflect the changes in the Chinese language in the last century. They also represent different styles of Chinese writing, ranging over narration, exposition and poetry. Due to the diversity of the translators' backgrounds, some versions follow the language standards of mainland China, while other have more Taiwan or Hong Kong flavor. But they have one thing in common: they were all done very professionally, with great care put into every sentence. Therefore the sentences are usually well-formed. All this makes the Chinese translations of the Bible a high-quality and well-balanced corpus of the Chinese language.

To study the linguistic features of this text corpus, we have been analyzing its syntactic structures with a Chinese parser in the last few years. The result is a grammar that covers all the syntactic structures in this domain and a dictionary that contains all the words in this text corpus. A lot of effort has also been put into tree-pruning and tree selection so that the bad trees can be filtered out. Therefore we are able to parse most of the sentences in this corpus correctly and produce a complete treebank of all the Chinese translations.

The value of such a treebank in the study and search of the Bible is obvious. But it should also be a valuable resource for computational linguistic research outside the Bible domain. After all, it is a good representation of the syntactic structures of Chinese.

2 The Data Set

The text corpus for the treebank includes eight different versions of Chinese translations of the Bible, both the Old Testament and the New Testament. They are listed below in chronological order with their Chinese names, abbreviations, and years of publication:

- Chinese Union Version (和合本 CUV 1919)
- Lv Zhenzhong Version (吕振中译本 LZZ 1946)
- Sigao Bible (思高圣经 SGB 1968)
- Today's Chinese Version (现代中文译本 TCV 1979)
- Recovery Version (恢复本 RCV 1987)
- New Chinese Version (新译本 NCV 1992)
- Easy-to-Read Version (普通话译本 ERV 2005)
- Chinese Standard Bible (中文标准译本 CSB 2008)

All these versions are in vernacular Chinese (白话文) rather than classical Chinese (文言文), with CUV representing "early vernacular" (早期白话文) and the later versions representing contemporary Chinese. The texts are all in simplified Chinese. Those translations which were published in traditional Chinese were converted to simplified Chinese. For a linguistic comparison of those different versions, see Wu *et al* (2009).

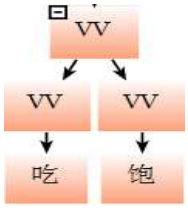
In terms of literary genre, more than 50% of the Bible is narration, about 15% poetry, 10% exposition, and the rest a mixture of narrative, prosaic and poetic writing. The average number of characters in a single version is close to one million and the total number of characters of these eight versions is 7,672,420.

Each book in the Bible consists of a number of chapters which in turn consist of a number of verses. A verse often corresponds to a sentence, but it may be composed of more than one sentence. On the other hand, some sentences may span multiple verses. To avoid the controversy in sentence segmentation, we preserved the verse structure, with one tree for each verse. The issues involved in this decision will be discussed later.

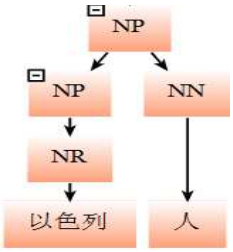
3 Linguistic Issues

In designing the tree structures, we essentially followed the Penn Chinese Treebank (PCTB) Guidelines (Xia 2000, Xue & Xia 2000) in segmentation, part-of-speech tagging and bracketing. The tag set conforms to this standard completely while the segmentation and bracketing have some exceptions.

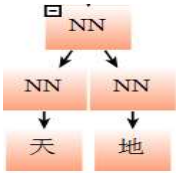
In segmentation, we provide word-internal structures in cases where there can be variations in the granularity of segmentation. For example, a verb-complement structure such as 吃饱 is represented as



so that it can be treated either as a single word or as two individual words according to the user's needs. A noun-suffix structure such as 以色列人 is represented as

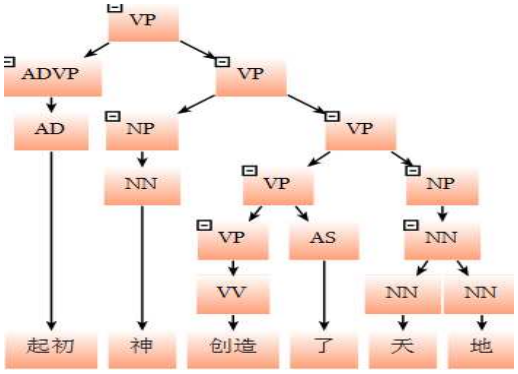


to accommodate the need of segmenting it into either a single word or two separate words. Likewise, a compound word like 天地 is represented as

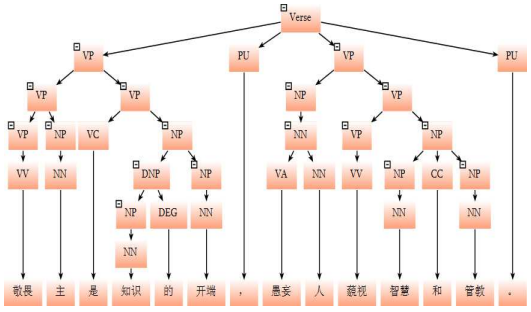


to account for the fact that it can also be analyzed as two words. This practice is applied to all the morphologically derived words discussed in Wu (2003), which include all linguistic units that function as single words syntactically but lexically contains two or more words. The nodes for such units all have (1) an attribute that specifies the word type (e.g. Noun-Suffix, Verb-Result, etc.) and (2) the sub-units that make up the whole word. The user can use the word type and the layered structures to take different cuts of the trees to get the segmentation they desire.

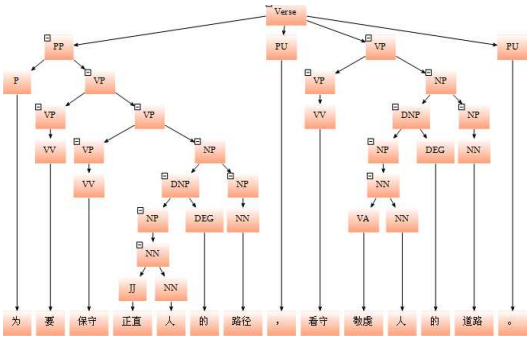
In bracketing, we follow the guidelines of Penn Chinese treebank, but we simplified the sentence structure by omitting the CP and IP nodes. Instead, we use VP for any verbal unit, whether it is a complete sentence or not. Here is an example where the tree is basically a projection of the verb, with all other elements being the arguments or adjuncts of the VP:



There are two reasons for doing this. First of all, we choose not to represent movement relationships with traces and empty categories which are not theory-neutral. They add complexities to automatic parsing, making it slower and more prone to errors. Secondly, as we mentioned earlier, the linguistic units we parse are verses which are not always a sentence with an IP and a CP. Therefore we have to remain flexible and be able to handle multiple sentences, partial sentences, or any fragments of a sentence. The use of VP as the maximal project enables us to be consistent across different chunks. Here is a verse with two sentences:

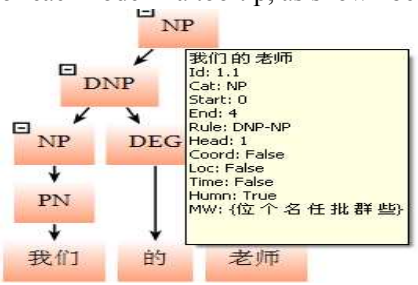


Notice that both sentences are analyzed as VPs and the punctuation marks are left out on their own. Here is a verse with a partial sentence:



This verse contains a PP and a VP. Since it is not always possible to get a complete sentence within a verse, we aim at "maximal parses" instead of complete parses, doing as much as we can be sure of and leaving the rest for future processing. To avoid the clause-level ambiguities as to how a clause is related to another, we also choose to leave the clausal relations unspecified. Therefore, we can say that the biggest linguistic units in our trees are clauses rather than sentences. In cases where verses consist of noun phrases or prepositional phrases, the top units we get can be NPs or PPs. In short, the structures are very flexible and partial analysis is accepted where complete analysis is not available.

While the syntactic structure in this treebank is underspecified compared to the Penn Chinese Treebank, the lexical information contained in the trees are considerably richer. The trees are coded in XML where each node is a complex attribute-value matrix. The trees we have seen above are visualizations of the XML in a tree viewer where we can also view the attributes of each node in a tooltip, as shown below:



Here, the attributes tell us among other things that (1) this node is formed by the rule "DNP-NP", (2) the head of this phrase is its second child (position is 0-based), (3) there is no coordination in this phrase, (4) this is not a location phrase, (5) this is not a time phrase, (6) the NP is a human being, and (7) the head noun can take any of those measure words (量词): 位，个，名，任，批，群 and 些. There are many other attributes and a filter is applied to determine which attributes will show up when the XML is generated.

4 Computational Issues

As we have mentioned above, the trees are generated automatically by a Chinese parser. It is well-known that the state-of–the-art natural language parsers are not yet able to produce syntactic analysis that is 100% correct. As a result, the automatically generated trees contain errors and manual checking is necessary. The question is what we should do when errors are found.

The approach adopted by most treebanking projects is manual correction which involves editing the tree files. Once the trees have been modified by hand, the treebank becomes static. Any improvement or update on the treebank will require manual work from then on and automatic parsing is out of the picture. This has several disadvantages. First of all, it is very labor-intensive and not everyone can afford to do it. Secondly, the corrections are usually token-based rather than type-based, which requires repetitions of the same correction and opens doors to inconsistency. Finally, this approach is not feasible with trees with complex feature structures where manual editing is difficult if not impossible.

To avoid these problems, we adopted the approach of dynamic treebanking (Oepen *et al* 2002) where corrections/updates are not made in the tree files but in the grammar and dictionary that is used to generate the trees. Instead of fixing the trees themselves, we improve the tree-generator and make it produce the correct trees. Every error found the trees can be traced back to some problem in the grammar rules, dictionary entries, or the tree selection process. Once a "bug" is resolved, all problems of the same kind will be resolved throughout the whole treebank. In this approach, we never have to maintain a static set of trees. We can generate the trees at any time with any kind of customization based on users' requirement.

Dynamic treebanking requires a high-accuracy syntactic parser which is not easy to build. A Chinese parser has the additional challenge of word segmentation and name entity recognition. These problems become more manageable once the texts to be parsed are narrowed down to a specific domain, in our case the domain of Biblical texts.

The dictionary used by our parser is based on the Grammatical Knowledge Base of Contemporary Chinese (GKBCC) licensed from Beijing University. It is a wide-coverage, feature-rich dictionary containing more than 80,000 words. On top of that, we added all the words in the eight translations, including all the proper names, which are not in the GKBCC. The total vocabulary is about 110,000 words. Since we follow the PCTB guidelines in our syntactic analysis, the grammatical categories of GKBCC were converted to the PCTB POS tags.

With all the words in the dictionary, which eliminates the OOV problem, the only problem left in word segmentation is the resolution of combinational ambiguities and overlapping ambiguities. We resolve these ambiguities in the parsing process rather than use a separate word segmenter, because most wrong segmentations can be ruled out in the course of syntactic analysis (Wu and Jiang 1998).

Our grammar is in the HPSG framework. In addition to feature-rich lexical projections, it also bases its grammatical decisions on the words in the preceding and following contexts. Multiple trees are generated and sorted according to structural properties. The treebank contains the best parse of each verse by default, but it can also provide the top *N* trees. The grammar is not intended to be domain-specific. Almost all the rules there apply to other domains as well. But the grammar is "domain-complete" in the sense that all the grammatical phenomena that occur in this domain are covered.

The developers of the treebank only look at the top tree of each verse. If it is found to be incorrect, they can fix it by (1) refining the conditions of the grammar rules, (2) correcting or adding attribute values in the lexicon, or (3) fine-tuning tree ranking and tree selection. For phrases which occur frequently in the text or phrases which are hard to analyze, we store their correct analysis in a database so that they can be looked up just like a dictionary entry. These "pre-generated" chunks are guaranteed to have the correct analysis and they greatly reduce the complexity of sentence analysis.

The same grammar and dictionary are used to parse the eight different versions. The development work is mainly based on CSB. Therefore the trees of the CSB text have higher accuracy than those of other versions. However,

due to the fact that all the eight versions are translations of the same source text, they share a large number of common phrases. As our daily regression tests show, most fixes made in CSB also benefit the analysis of other versions.

5 Evaluation

Due to the optimization of the grammar and dictionary for the Bible domain, the accuracy of this Chinese parser is much higher than any other general-purpose Chinese parsers when the texts to be parsed are Chinese Bible texts. Therefore the accuracy of the trees is higher than any other automatically generated trees. Unfortunately, there is not an existing treebank of Chinese Bible translations that can be used as a gold standard for automatic evaluation. We can only examine the quality through manual inspection. However, there does exist a segmented text of the CUV translation.[1] Using this text as the gold standard is ideal because the development data for our system is CSB rather than CUV or other versions.

As we have mentioned above, the segmentation from the trees can be customized by taking different cuts in cases where word-internal structures are available. In order to make our segmentation match the existing CUV segmentation as closely as possible, we studied the CUV segments and made a decision for each type of words. For example, in a verb-complement construction where both the verb and the directional/resultative complement are single characters, the construction will be treated as a single word.

We evaluated the segmentation of our CUV trees with the scoring script used in the first international Chinese segmentation bakeoff (Sproat & Emerson 2003). Here are the results:

Recall:	99.844%
Precision:	99.826%
F-Score:	99.845%

We don't show the OOV numbers as they are not relevant here, because all the words have been exhaustively listed in our dictionary.

Of a total of 31151 verses in the Bible, 30568 verses (98.13%) do not contain a single error (whole verses segmented correctly).
Of course, segmentation accuracy does not imply parsing accuracy, though wrong segmentation necessarily implies a wrong parse. Since we do not have a separate word segmenter and segmentation is an output of the parsing process, the high segmentation accuracy does serve as a reflection of the quality of the trees. There would be many more segmentation errors if the trees had many errors.

6 Use of the Treebank

The treebank has been used in the area of Bible translation and Bible search. In Bible translation, the trees are aligned to the trees of the original Hebrew and Greek texts[2]. By examining the correspondences between the Chinese trees and the Hebrew/Greek trees, one is able to measure how faithful each translation is to the original. In Bible search, the trees makes it possible to use more intelligent queries based not only on words but on syntactic relations between words as well.

An obvious use of the treebank is to train a statistical parser. Though the domain speci-

[1] The segmented CUV text was provided by Asia Bible Society.

[2] The Hebrew and Greek trees were also provided by Asia Bible Society.

ficity of the treebank makes it less likely to build from it a good lexicalized statistical parser that can be used in the general domain, we can still extract a lot of non-lexical syntactic information from it. It can fill many of the gaps in the parsers that are built from other treebanks which consist mainly of news articles.

A special feature of this treebank is that it is built from a number of parallel texts -- different Chinese translations of the same verses. By aligning the parallel trees (ideally through the original Hebrew and Greek trees as pivots), we can acquire a knowledge base of Chinese synonyms and paraphrases. Presumably, the different Chinese subtrees corresponding to the same Hebrew/Greek subtree are supposed to convey the same meaning. The words and phrases covered by those subtrees therefore represent Chinese expressions that are synonymous. A knowledge base of this kind can be a valuable addition to the lexical study of Chinese.

7 Summary

We presented a Chinese treebank of parallel Bible translations. The treebank is built through dynamic treebanking where the trees are automatically generated by a Chinese parser optimized for parsing Biblical texts. The trees can serve as a useful resource for different language projects.

References

Sproat, Richard and Thomas Emerson. 2003. The First International Chinese Segmentation Bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, July 11-12, Sapporo, Japan.

Wu, Andi, J. and Z. Jiang, 1998. Word segmentation in sentence analysis, in Proceedings of 1998 International Conference on Chinese Information Processing, pp. 46--51. 169--180, Beijing, China.

Wu, Andi. 2003. Customizable Segmentation of Morphological Derived Words in Chinese. In *International Journal of Computational Linguistics and Chinese Language Processing*, 8(1):1-27.

Wu, And, Arron Ma, Dong Wang. 2009. Fidelity and Readablity – a quanatative comparison of Chinese translations of the New Testament. Proceedings of the Conference on "Interpretation of Biblical Texts in Chinese Contexts", Sichuan Univeristy, December 2009.

Xia, Fei. 2000. *Segmentation Guidelines for the Penn Chinese Treebank (3.0)*. Technical Report, University of Pennsylvania.

Xia, Fei. 2000. *The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0)*. Technical Report, University of Pennsylvania.

Xue, Nianwen and Fei Xia. 2000. *The Bracketing Guidelines for the Penn Chinese Treebank (3.0)*. Technical Report, University of Pennsylvania.

Oepen, Stephan, Dan Flickinger, Kristina Toutanova, Christoper D. Manning. 2002. LinGO Redwoods: A Rich and Dynamic Treebank for HPSG In *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT2002)*, Sozopol, Bulgaria.