GBI Treebanks as a Resource for New Applications

Andi Wu
Global Bible Initiative
andi.wu@globalbibleinitiative.org

Abstract

Global Bible Initiative (GBI) have developed Hebrew OT treebanks and Greek NT syntactic treebanks. The treebanks were first generated with a parser using computerized Hebrew and Greek grammars and then proofed verse by verse by Hebrew and Greek Scholars. All the corrections made by the scholars were kept as disambiguation data.

The phrase structures in the trees have been used to build interlinears, concordances, and translation memories which operate not only on the word level, but on the phrase and clause levels as well. The syntactic relations (dependencies) in the trees have also been used to do smart search where we can find texts that are different in form but similar in meaning.

Recently, we have also used the trees to improve the accuracy of automatic word alignment and explore tree-based interactive machine translation of the Bible. The auto aligner can be used to the Hebrew and Greek texts to translations in various languages. The interactive machine translation will speed up Bible translation without compromising quality by providing real time suggestions and checking.

We have already contributed two sets of Greek trees to Creative Commons, the Nestle 1904 version and the SBLGNT version. We also have trees for NA27 and NA28, but we do not own the texts. The Hebrew OT treebank we developed was owned by the Groves Center. We are also capable of creating new treebanks with the parser, grammar, and disambiguation data we own if we are given a text that is morphologically tagged.

Creation of Treebanks

Since 2005, Global Bible Initiative (GBI) has been creating treebanks of Biblical texts, both the original Hebrew and Greek texts, and their translations in English and Chinese. The trees are all in XML, and were machine-generated with human guidance. The actual steps of creation are as follows.

(1) A syntactic parser was developed. The parser can work with the dictionary and grammar of any language.
(2) A dictionary and a computerized grammar was developed for each language. In the case of Hebrew OT or Greek NT, the dictionary was created from a morphologically tagged text, and the grammar was hand-crafted over a long period of time through many cycles of development and testing.
(3) The texts were parsed using the dictionary and grammar, resulting in the initial treebank.
(4) The trees were checked and corrected verse by verse by Hebrew and Greek scholars. The corrections were digitally captured during the process and stored in a database which was later used to guide the parser in cases of ambiguity.
(5) The texts were parsed again using the dictionary, grammar, plus the disambiguation database. Whenever there is more than one choice in the parsing process, the parser will follow the paths set by the scholars.

(4) and (5) went through many iterations until the scholars were happy with the results.

The above process has been used to complete the following original language treebanks:

- Hebrew OT treebank based on the MORPH 4.20 database of the Groves Center of Advanced Biblical Research (formerly the Westminster Hebrew Institute).
- Hebrew OT prosodic treebank where the structure is based on the accents/cantillation marks in MORPH 4.20.
- 5 Greek NT treebanks:
    - NA28 text with GBI's morphology
    - NA27 text with James Tauber's morphology
    - Nestle 1904 text with Urik Peterson's morphology
    - SBLGNT with James Tauber's morphology
    - SBLGNT with Logos morphology

We also created English treebanks for HCSB, NASB, ESV and NRSV, and Chinese treebanks for CUV and CSB.

Applications of Treebanks

In GBI's Bible translation projects, the trees have helped translators to get a clearer view of sentence structures, so that even people not fluent in Hebrew and Greek can have a better understanding of what the original text is saying.

The phrase structures in the trees have been used to build tree-based translation memories, concordances, interlinears and reverse interlinears, where the linguistic units are not limited to single words, as is traditionally the case, but contain higher-level constituents and long-distance dependencies as well. Translators can check consistency not only at the level of words, but also at the levels of phrases, clauses and collocations. This makes everything more context-sensitive, which helps to filter out irrelevant cases and narrow down the checking space.

The syntactic relations in the trees (dependencies), together with the probabilistic synonym database we developed (which will not be discussed in this paper), have enabled us to conduct smart search in the Bible. Traditional Bible search is only able to find texts that contain words that are identical to the words in the query. Our search methods can also find texts that have similar meanings but contain different words in different orders. Meaning is defined as "similar words in similar syntactic relations". We therefore search on meaning rather than on form. The user can pick a phrase or clause in the text and find other phrases and clauses that are related in meaning but may look very different.

In the realm of language learning, the treebank can help us find sentences that contain identical or similar grammatical structures. A statistical study of the phrases and clauses can also help us identify high-impact chunks of text which should be prioritized in the curriculum. In our translation projects, these high impact chunks have made a difference in cases where prioritization is required, such as key-phrase checking.

Recently, we have also been exploring the use of trees in automatic word alignment and interactive machine translation. The two projects will be described in more detail below.

Automatic Word Alignment

Alignment is the process whereby the words in the original texts (source texts) are linked to the corresponding words in a translation (target text). This is the foundation for key-term checking, translation memories, interlinears, and ultimately machine translation and automatic evaluation of translations. Manual alignment is not feasible because it may take years to complete a single version of translation.

Statistical word aligners are publicly available, and have been widely used in machine translation. However, their performance is not satisfactory in the Bible domain, where higher accuracy is required while the size of the data is not big enough for training a high-accuracy aligner. In addition, the statistical aligners are not sensitive to syntactic structure. They tend to make more mistakes when the same word occurs multiple times in a sentence while the syntactic structure of the target sentence is very different.

In our experiment, we integrated the statistical models into a tree-based decoding algorithm. We traverse the tree bottom-up, keeping the top N candidates at each level, and pick the top candidate when the root node is reached. In other words, the alignment is done segment by segment and layer by layer, which can tolerate more variation in syntactic structure. Words that are not likely to form a single syntactic unit in the target sentence (e.g. words are far apart from each other) will not be aligned to a single unit in the source sentence.

We tested this tree-based aligner on 18 different translations in 7 different languages. Compared with the results from a pure statistical parser, the tree-based aligner is able to reduce the error rate by 40-50%. The reduction of error rate is especially obvious in the alignment of content words which are more important for other applications.

This new aligner can be used to align a translation in any language if the text is already word-segmented, i.e. with spaces between words. (Agglutinative languages are still a challenge.) Once the words are aligned, we can align phrases and clauses on the basis of the Hebrew and Greek trees. This will enable us to build translation memories, concordances and interlinears for that translation. The alignment can also be used to gloss the original texts in that language, which can facilitate Bible translation and language learning.

Interactive Machine Translation

In interactive machine translation, the translator is not given a machine translated draft to edit. Instead, the machine observes what the translator has typed and provide suggestions which the translator can either accept or reject. The suggestions are based on the translation memory built from the texts that have already been previously, including the ones being currently translated. The system learns in real-time and allows the system to progressively provide better suggestions to human translators. It can thus speed up Bible translation without compromising quality.

As the translator types along, good suggestions are expected to pop up at the right time. However, without any knowledge of the structure of a sentence, it is difficult for the machine to know what to suggest at a particular point in the sentence being translated. The suggestion is to be prompted by

what the translator has typed the so far, especially the last few words. If the target language has a word order that is very similar to the source language, this will not be a big problem. We can simply show a possible translation of the next word in the source text. This will not work if word orders are very different.

This is where the trees can help. In the experimental system we are building, we are translating the tree of a text. A tree-based translation memory is built along the way. The last N words typed by the translator (the suffix) are dynamically linked to words in the tree. If the suffix can be linked to part of a given subtree and this subtree can be found in the existing translation memory, the machine will make suggestions. It will suggest how to complete the translation of the current subtree only. The suggestions are therefore position-sensitive and constrained by local contexts. As a result, they are always relevant to the focus point of the translator. From the viewpoint of the source text, the translation does not go linearly from left to right or right to left, but go subtree by subtree. It can start from the beginning, the middle, or the end of the tree. The machine will not try to suggest translations for a parent node until at least one of its child nodes has been translated. The order in which the nodes in the tree are visited can be in any order depending on the word order of the target language.

Interactive machine translation can also offer real-time checking. Since the translation is constantly being linked to a subtree, we can always check if this subtree exists in the translation memory (i.e. if this subtree occurs elsewhere and if it has been previously translated) and see if the current translation is different from the one(s) in the translation memory. It is common for a single word to have different translations in different contexts, but if a multiple word phrase has different translations in different places, a flag should be raised.

We have not tried using trees in the realm of language learning, but it is not hard to come up with some use cases. When a new phrase is being taught, for example, we would like to know if this phrase occurs anywhere else. Teachers can use this information to prepare teaching materials and students can use it for practice and review. Without the trees, we can find out whether a single word occurs elsewhere, but the list can be too long. We are probably only interested in the ones that occur in a given context. The trees provide those local contexts by grouping words that form phrases.

Availability of Treebanks

The GBI treebanks were developed along with the Bible translation projects that GBI has been undertaking. They were initially meant for internal use to support our translation efforts. They caught the attention of the outside world later and we realized that there is a great need for the trees in Bible translation and Bible engagement. To support worldwide evangelism, we are willing to share our resources to other workers in this field.

We have already put two sets of Greek NT trees in Creative Commons: the Nestle 1904 trees with Urik Peterson's morphology and the SBLGNT trees with James Tauber's morphology. We would like to share the NA27 and NA28 trees as well, but their texts are owned by German Bible Society, which prevents us from making the trees open source.

The Hebrew OT treebank were co-developed with Kirk Lowery of the Groves Center which owns the trees now. We did not ask for ownership when the collaboration started because we never expected to make commercial use of it. So we do not own the trees in spite of the fact that we did 95% of the work in creating the treebank. The agreement between GBI and Groves Center allow us to use the trees commercially only in the non-English-speaking world, such as China. We wish this treebank could be put in the public domain, and we are still negotiating with Groves Center on this. They are open to the idea of open source if there is enough financial incentive.

GBI does own the technology that was used the create the Hebrew OT treebank, which includes the parser, the grammar, and the tree editor, and the disambiguation data created by our Hebrew Scholars during the tree-building process. If we are given another morphologically tagged text, we can create another dictionary and use that to generate another set of trees. We will put this new treebank in Creative Commons.

After completing the treebanks, we have also been adding new data to the trees, which include:

- Pronominal reference. For each pronoun and implied subject in subject-less clauses, we specify the antecedent(s) the pronoun or implied pronoun refers to.
- Verbal frame. For each verb, we specify its argument structure or valency.
- Semantic Roles. For each argument in the verb frame, we specify its role in the clause, such as agent, patient, time, location, instrument, etc.
- Strong numbers and Extended Strong numbers. The latter was developed by GBI. It is more fine-grained and accurate.
- Sentence-based trees in addition to verse-based trees (completed for Greek NT only).
- English and Chinese glosses (still being refined).
- Greek equivalents of Hebrew words and Hebrew equivalents of Greek words (need to be manually checked)
- Word senses. For words that have more than one sense, each occurrence of the word is marked with a sense number (need to be manually checked).
- Logical forms of sentences (just started)

In the case of Hebrew OT, these additional data are not owned by the Groves Center, since they are not part of the initial agreement.

Unlike the Hebrew OT syntactic treebank, the Hebrew OT prosodic treebank is not owned by the Groves Center, though it owns the text. The creation of this treebank does not require a morphology, as it depends on the cantillation marks in the text only.

Conclusion

GBI has created a set of original language treebanks that have resulted in innovative applications in the Bible translation and Bible Engagement. We want to let other people in these fields benefit from the work we have done by making the data available and sharing our experience in using the data. There are still some legal, technical, and financial obstacles in the process of making the data public. Hopefully these obstacles can be removed soon.