Incorporating Alternate Translations into English Translation Treebank

Ann Bies, Justin Mott, Seth Kulick, Jennifer Garland, Colin Warner
Linguistic Data Consortium
University of Pennsylvania
3600 Market Street, Suite 810
Philadelphia, PA 19104 USA
E-mail: {bies,jmott,skulick,garjen,colinw}@ldc.upenn.edu

Abstract

New annotation guidelines and new processing methods were developed to accommodate English treebank annotation of a parallel English/Chinese corpus of web data that includes alternate English translations (one fluent, one literal) of expressions that are idiomatic in the Chinese source. In previous machine translation programs, alternate translations of idiomatic expressions had been present in untreebanked data only, but due to the high frequency of such expressions in informal genres such as discussion forums, machine translation system developers requested that alternatives be added to the treebanked data as well. In consultation with machine translation researchers, we chose a pragmatic approach of syntactically annotating only the fluent translation, while retaining the alternate literal translation as a segregated node in the tree. Since the literal translation alternates are often incompatible with English syntax, this approach allows us to create fluent trees without losing information. This resource is expected to support machine translation efforts, and the flexibility provided by the alternate translations is an enhancement to the treebank for this purpose.

Keywords: English Translation Treebank, alternate translations, parsing, idiomatic expressions

1. Introduction

The English Treebank team at the Linguistic Data Consortium has previously responded to challenges from informal web genres with the annotation of corpora such as the English Web Treebank (Bies, et al. 2012). However, a new challenge arose with a parallel English/Chinese corpus of web data that includes alternate English translations (one fluent, one literal) of expressions that are idiomatic in the Chinese source.

New annotation guidelines and new processing methods were both required to account for the alternates. In consultation with machine translation researchers, we chose a pragmatic approach of syntactically annotating only the fluent translation, while retaining the alternate literal translation as a segregated node in the tree. Since the literal alternates are often incompatible with English syntax, this approach allows us to create fluent trees by segregating the literal alternates under easily identifiable tree nodes.

2. Why Include Alternate Translations in the Annotated Data?

In previous machine translation (MT) programs such as GALE, alternate translations of idiomatic expressions had been present in MT evaluation data only, but due to the high frequency of such expressions in informal genres such as discussion forums, MT system developers in the BOLT program requested that alternatives be added to the training data as well. For example, the English translation *Teacher salaries are very high, is that so? If there comes a day when everyone will become a teacher [by hook or by crook | sharpen their heads], as popular as the civil servants of today, that will mean that teachers are really attractive* contains both a fluent translation (*by hook or by crook*) and the literal translation (*sharpen their heads*) for the idiomatic expression in the original Chinese (削尖脑袋 xuè jiān nǎodai). The purpose of alternative translations in the evaluation data was to accommodate the needs of monolingual English-speaking annotators performing human-mediated translation error rate (HTER) evaluation tasks (Snover, et al. 2006). In HTER scoring, a monolingual English-speaking annotator is presented with MT output and a corresponding gold standard human reference translation. The task of the annotator is to edit the MT output to match the meaning of the reference translation, keeping the number of edits to the minimum required to achieve semantic equivalence. When the source data contains idiomatic expressions whose intended meaning is not easily understood from a literal translation, such as the example above, the reference translation is augmented with alternatives which include both the intended and literal meanings. This allows the HTER annotator to give MT systems credit for matching the literal meaning of the expression. In the BOLT program, system developers requested access to these alternatives in the training data as well, in order to facilitate better handling of idiomatic expressions by their systems. As a result, these alternates are present for the first time in the English data available for treebank annotation.

3. Adapting the Parsing Process to Account for Translation Alternates

The parsing process was modified to account for the alternate translations[1]. The input annotation for parsing

[1] The parser used in this work was the Bikel parser (www.cis.upenn.edu/~dbikel/software.html). However, the

included, as usual, word-level tokenization and part-of-speech (POS) annotation, and for this translated data the input included both the literal and fluent translation alternates. In addition, pre-annotated tree nodes were inserted prior to parsing to indicate the scope of the alternates. This was done mainly automatically with subsequent manual validation. Variation in the mark-ups for the alternate translations required some manual insertion as well. The syntactic node META was used to indicate the full extent of the literal translation alternate, and also for the metadata punctuation delimiters of the literal translation alternate (the open square bracket preceding the fluent translation alternate, the pipe preceding the literal alternate, and the close square bracket following it). Metadata punctuation tokens delimiting the alternate translations also received the usual POS tags for these punctuation marks[2]: (-LRB- [), (-RRB-]), and (SYM |).

For example, this input to the parsing pipeline indicates that *by hook or by crook* is the fluent translation to the parsed and treebanked, while *sharpen their heads* is the literal translation:

```
...everyone/NN will/MD become/VB a/DT
teacher/NN (META [/-LRB-) by/IN hook/NN
or/CC by/IN crook/NN (META |/SYM
sharpen/VB their/PRP$ heads/NNS) …
```

The first step of the parsing process then simply removed all the tokens inside the META pre-bracketing to create the input to the parser. After the parser produced a tree, the sections of META information were re-inserted, with the tokens *sharpen their heads* left flat under the META node. The resulting tree was then used in the annotation process, as input to manual syntactic annotation and correction. One final point is that this process of using the annotated META information served as a check on the manual clean-up and post-processing steps mentioned in footnote 2.

4. New Treebank Guidelines to Account for Translation Alternates

Because both the literal and the fluent alternates[3] were present in the text of English data available for annotation, it was necessary to account for them both in the tree. However, it was not feasible to syntactically annotate the literal alternates in the context of the full tree, since they often do not make sense as part of the English syntax. For example, if we use only the literal alternate in this case, the resulting clause (*If there comes a day when everyone will become a teacher sharpen their heads*) is not annotatable as an English clause.

We adopted a pragmatic approach to creating guidelines for annotating this data that allowed both alternates to appear in the tree, but also allowed the fluent alternate to be the base of the syntactic annotation.

1. Both literal and fluent translation alternates were annotated for word-level tokenization and part-of-speech. For example, all syntactic and POS nodes are shown here:

```
(PP-MNR (META (-LRB- [))
        (PP (IN by)
            (NP (NN hook)))
        (CC or)
        (PP (IN by)
            (NP (NN crook)))
        (META (SYM |)
              (VB sharpen)
              (PRP$ their)
              (NNS heads)
              (-RRB- ]))
```

2. Only the fluent translation alternates were annotated as part of the syntactic structure of the tree. Syntactic structure was not annotated inside the META node, since the syntax of the literal alternates often does not fit syntactically into the surrounding tree. The META note was attached at the level of the fluent translation, or as close to it as possible. This was done without adding additional structure to the tree, so that, upon removal of the META node, a valid tree remained. Annotators were instructed to ignore the literal translations in order to assign the most fluent analysis possible.

For example, in the tree below, the literal alternate was included under the META node, but only the fluent alternate was syntactically annotated:

parser itself was not modified for this work, and the changes described here were a wrapper around the parser.

[2] After part-of speech annotation, the literal translations were delimited automatically by script and placed under a treebank-like META node. Variation in the notation to mark off the alternates necessitated some manual clean up and post processing to achieve this. In most cases, the literal translations were contained within the span from | to]. In a small number of cases there was some variation in the markup that is delimiting the translation alternates. For example, l appears in place of the expected |, } for] , the initial [bracket may be missing, and the final] bracket may be missing. For treebank purposes, we marked the actual translation alternates and the existing markup with META nodes, regardless of such variation.

[3] One literal translation alternate and one fluent translation alternate were provided for each idiomatic phrase.

If there comes a day when everyone will become a teacher [by hook or by crook | sharpen their heads], as popular as the civil servants of today, that will mean that teachers are really attractive.[4]

```
(S (SBAR-ADV If
         (S (NP-SBJ there)
            (VP comes
                (NP (NP a day)
                    (SBAR (WHADVP-1 when)
                          (S (NP-SBJ everyone)
                             (VP will
                                 (VP become
                                     (NP-PRD (NP a teacher)
                                             (ADJP-2 *ICH*))
                                     (PP-MNR (META -LRB- )
                                             (PP by
                                                 (NP hook))
                                            or
                                             (PP by
                                                 (NP crook))
                                             (META | sharpen their heads -RRB- ))
                                     ,
                                     (ADJP-2 (ADJP as popular)
                                             (PP as
                                                 (NP (NP the civil servants)
                                                     (PP of
                                                         (NP today)))))
                                     (ADVP-TMP-1 *T*)))))))))
    ,
    (NP-SBJ that)
    (VP will
        (VP mean
            (SBAR that
                  (S (NP-SBJ teachers)
                     (VP are
                         (ADJP-PRD really attractive))))))
    .)
```

3. A version of the trees was generated with the META nodes and their children removed from the trees, so only the fluent alternates appear in the resulting trees. This was an entirely automatic post-processing step following annotation. For example, in the tree below, only the fluent alternate remains and the literal alternate is removed, creating a fluent tree:

[4] The trees for this example are simplified here by removing the part-of-speech nodes, for readability, and are shortened to include only the sentence that includes the translation alternate. They are also indented to make the tree structure more visible. However, for completeness, the full tree for this SU is as follows in the release format:

```
( (SQ (S (NP-SBJ (NN Teacher) (NNS salaries)) (VP (VBP are) (ADJP-PRD (RB very) (JJ high)))) (, ,) (SQ
(VBZ is) (NP-SBJ (DT that)) (ADVP-PRD (RB so))) (. ?)) (S (SBAR-ADV (IN If) (S (NP-SBJ (EX there)) (VP
(VBZ comes) (NP (NP (DT a) (NN day)) (SBAR (WHADVP-1 (WRB when)) (S (NP-SBJ (NN everyone)) (VP (MD will)
(VP (VB become) (NP-PRD (NP (DT a) (NN teacher)) (ADJP-2 (-NONE- *ICH*))) (PP-MNR (META (-LRB- [)) (PP
(IN by) (NP (NN hook))) (CC or) (PP (IN by) (NP (NN crook))) (META (SYM |) (VB sharpen) (PRP$ their) (NNS
heads) (-RRB- ])))) (, ,) (ADJP-2 (ADJP (SYM =) (RB as) (JJ popular)) (PP (IN as) (NP (NP (DT the) (JJ
civil) (NNS servants)) (PP (IN of) (NP (NN today)))))) (ADVP-TMP-1 (-NONE- *T*))))))))))) (, ,) (NP-SBJ
(DT that)) (VP (MD will) (VP (VB mean) (SBAR (IN that) (S (NP-SBJ (NNS teachers)) (VP (VBP are) (ADJP-PRD
(RB really) (JJ attractive))))))) (. .)) )
```

If there comes a day when everyone will become a teacher by hook or by crook, as popular as the civil servants of today, that will mean that teachers are really attractive.

```
(S (SBAR-ADV If
        (S (NP-SBJ there)
           (VP comes
               (NP (NP a day)
                   (SBAR (WHADVP-1 when)
                         (S (NP-SBJ everyone)
                            (VP will
                                (VP become
                                    (NP-PRD (NP a teacher)
                                            (ADJP-2 *ICH*))
                                    (PP-MNR (PP by
                                                (NP hook))
                                            or
                                            (PP by
                                                (NP crook)))
                                    ,
                                    (ADJP-2 (ADJP as popular)
                                            (PP as
                                                (NP (NP the civil servants)
                                                    (PP of
                                                        (NP today)))))
                                    (ADVP-TMP-1 *T*)))))))))
   ,
   (NP-SBJ that)
   (VP will
       (VP mean
           (SBAR that
                 (S (NP-SBJ teachers)
                    (VP  are
                        (ADJP-PRD really attractive))))))
   .)
```

Both versions of the trees were published in the corpus: a version with both literal and fluent translation alternates included, and also a version with the literal alternates removed.

5. Distribution of META Nodes in the Annotated Corpus

After the treebank annotation was completed, it was possible to examine the distribution of the META nodes in this corpus. The data contained 577 META nodes accounting for 288 alternate translations affecting 248 sentences (out of a total of 5012 sentences). The most common category containing an alternate translation was NP (136 instances), followed by VP (88 instances) and S (23 instances). Other nodes affected were: ADJP, ADVP, FRAG, INTJ, NML, PP and UCP, each with less than 10 instances.

In 119 instances overall the literal translation maps exactly to the span of a single syntactic node. That is, there is the following structure:

```
(XP (META [)
    fluent translation
    (META | literal translation ]))
```

Again, most of these cases were NP (43), followed by VP (41) and S (12). In most of these cases the alternate translations are of the same category as the fluent translation. Examples of the three most common categories are below.

- NP: the literal translation maps exactly to the span of a noun phrase in the fluent translation and tree

[Anything | Divine horse] is possible.

```
(S (NP-SBJ (META [)
           Anything
           (META | Divine horse ]))
    (VP is
        (ADJP-PRD possible))
   .)
```

- VP: the literal translation maps exactly to the span of a verb phrase in the fluent translation and tree

The government of Fuyu County [is so selfish and calculating | plays the little abacus so well] !!

```
(S (NP-SBJ (NP The government)
           (PP of
               (NP Fuyu County)))
   (VP (META [)
       is
       (ADJP-PRD so selfish
                 and calculating)
       (META | plays the little
             abacus so well ]))
   !!)
```

Or should we [bury our head in the sand | plug our ears while stealing a bell] and continue with our harmonious dream?

```
(SQ Or
    should
    (NP-SBJ we)
    (VP (VP (META [)
            bury
            (NP our head)
            (PP-LOC in
                    (NP the sand))
            (META | plug our ears while
                  stealing a bell ]))
        and
        (VP continue
            (PP-CLR with
                    (NP our harmonious
                        dream))))
    ?)
```

- S: the literal translation maps exactly to the span of a full sentence in the fluent translation and tree

[By looking at small details, we can see big problems | By seeing only one spot of the leopard through a tube, we can get an overall idea about it as a whole].

```
(S (META [)
   (PP-MNR By
           (S-NOM (NP-SBJ *PRO*)
                  (VP looking
                      (PP-CLR at
                          (NP small
                          details)))))
   ,
   (NP-SBJ we)
   (VP can
       (VP see
           (NP big problems)))
   (META | By seeing only one spot of the
         leopard through a tube, we can
         get an overall idea about it as
         a whole ])
   .)
```

In the cases where the alternate translation does not map exactly to a syntactic node, there are several common patterns. In many cases, the alternate translation corresponds to a subspan of the node. For NPs the translation alternates often share a determiner or other pre-modifier:

> *But the [Russians | Old Hairy Ones] occupied it with no sign of letting go. If you have the ability, come and grab it!*
>
> *dear readers and the [original poster | floor host] do n't need to cherish much hope for this thing, and do n't count on it being solved fairly and justly.*

Also common in NPs are instances where the alternates are single token (for the fluent alternate) pre-modifiers:

> *Thus, no matter which way you look at it, a [democratic | MZ] world is beneficial and innocuous for Americans.*

Finally, there are a number of cases where, due to the surrounding structure, the literal translations and fluent translations are at different levels in the tree, so cannot be annotated as sisters. In the following tree, for example, the PP node *of GDP* prevents the META node dominating the literal alternate *red lights all the way to* from being annotated as a sister to the fluent NP *unqualified figures* (or as a sister to the full NP *unqualified figures of GDP*):

Those who have bad luck will face [unqualified figures of | red lights all the way to] GDP.

```
(S (NP-SBJ (NP Those)
           (SBAR (WHNP-9 who)
                 (S (NP-SBJ-9 *T*)
                    (VP have
                        (NP bad
                            luck)))))
    (VP will
        (VP face
            (NP (NP (META  [)
                    unqualified
                    figures)
                (PP of
                    (META | red lights
                          all the way
                          to ])
                    (NP GDP))))
    .)
```

6. Conclusion

A total of 147,433 tokens/words (145,427 tokens after translation alternates are removed) of English/Chinese Treebank web data (discussion forum genre) has now been treebanked and released as e-corpora (Bies, et al. 2013a; Bies, et al. 2013b). This data is consistent with the most current updated English Treebank Annotation Guidelines at LDC, and it will be published in the LDC Catalog in the near future.

In addition, this data has received multiple annotations. The Chinese Treebank of the source Chinese data and Chinese-English word alignments have been completed[5] for this data as well. These combined resources make a parallel aligned English-Chinese treebank of web data possible for the first time. This resource is expected to support machine translation efforts, and the flexibility provided by the alternate translations is an enhancement to the treebank for this purpose.

7. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-11-C-0145. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

8. References

Ann Bies, Justin Mott, Colin Warner, Seth Kulick. (2012). *English Web Treebank*. Linguistic Data Consortium, Catalog number LDC2012T13.

Ann Bies, Justin Mott, Colin Warner, Seth Kulick. (2013a). *BOLT Phase 1 English Treebank DF Part 6 V1.0 – ECTB*. Linguistic Data Consortium, Catalog number LDC2013E50.

Ann Bies, Justin Mott, Colin Warner, Seth Kulick. (2013b). *BOLT Phase 1 English Treebank DF Part 7 V1.0 – ECTB*. Linguistic Data Consortium, Catalog number LDC2013E76.

Daniel Bikel. (2004). *On the Parameter Space of Lexicalized Statistical Parsing Models*. Dissertation, Department of Computer and Information Sciences, University of Pennsylvania.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2006)*, Cambridge, Massachusetts.

[5] BOLT Phase 1 Chinese Treebank DF, Parts 1-4, LDC2012E109, LDC2012E120, LDC2012E130, LDC2013E32; BOLT Phase 1 Chinese Parallel Word Alignment and Tagging DF, Parts 1-5, LDC2012E24, LDC2012E72, LDC2012E95, LDC2013E02, LDC2013E51.