# Literature Review on Bitext Parsing

**Syntactic Parsing: Assignment 2**
**Shifei Chen**

---

## 1 Introduction

Natural Language comes with ambiguity. For example in English it is often the case that prepositional phrase (PP) can modify both the direct or distant preceding noun. Like the sentence "I booked a flight from LA", can either mean that "I" have booked a flight which will take off from LA, or "I" have booked a flight from (someone or some agency located in) LA.

But the interesting things, natural languages usually are not ambiguous in the same way (Huang et al., 2009) as their distant relatives. In other words, an ambiguity in one language often is clear in another unrelated, structurally differernt natural language. Like the sentence above, in Japanese one should explicitly add the particle "の" after "from LA" to distinguish these two semantics[1].

Therefore it is intuitive to leverage information from another language to help solve the ambiguity in the target language. I have choosen two papers from Burkett and Klein (2008) and Huang et al. (2009), which both presented encouraging results in this field. Both of the researches were focused on the Penn Chinese treebank and its English translations from Xue et al. (2002).

## 2 Measurements for the Bitext Trees

In order to leverage the information from one language to another, both papers agreed on an idea that each node in a sentence of one language should align to its couter part in the corrsponding translation, even though the span of the node may not be exactly the same.

### 2.1 The Model

Burkett and Klein (2008) designed their model as

$$P(t, a, t'|s, s') \tag{1}$$

which is the probability of the triple $(t, a, t')$. Here $t$ and $t'$ are the trees parsed from the source language and the target language, while $a$ is the alignment, or the at-most-one-to-one matchings between the tree pair. The euqation above is a general log-linear(maximum extropy) distribution over the triple $(t, a, t')$ for a given sentence pair $(s, s')$.

### 2.2 Alignments

They have also defined several features based on the theory to measure that alignment between two languages. For a node $n$ in the source language and the node $n'$ in the target language, they have defined $a(v, v')$ as the notation of the alignment for words inside the bispan, which is the posterior probability calculated by an independent word aligner. They believe that a good node alignment means that should be more word-to-word alignment for each word inside the nodes $n$ and $n\prime$. So for each node pair, $\sum_{v\in i(n)} \sum_{v'\in i(n')} a(v, v')$ means the sum of word alignments for each word inside the source node $n$ and the target node $n'$. $\sum_{v\in i(n)} \sum_{v'\in o(n')} a(v, v')$ and $\sum_{v\in o(n)} \sum_{v'\in i(n')} a(v, v')$ are the measurements to check the probability of a word alignment pair if one of hte word is outside the bispan. In the final experiments, they used a hard alignment feature (take the hard top-1 output from the aligner instead of all of it) and the scaled alignment feature (alignment divided by the size of the span) that were derived from these three base alignment features.

[1]The two sentence are "LA からのフライトを予約しました" and "LA からフライトを予約しました", which mean "I booked a flight that takes off from LA" and "I booked a flight from (someone or some agency located in) LA", respectively.

### 2.3 Other Features

Beside word alignemnt features, Burkett and Klein (2008) have also defined tree structure features and some monolingual features. In later experiments, they have showed that all bilingual features are proved to be better in contributing to the model's performance than monolingual features.

## 3 Training

During traininig Burkett and Klein (2008) faced the problem that the weights, which maximize the marginal log likelihood of what they did observe given their sentence pairs, were hard to compute. So they have developed several approximations and modifications.

### 3.1 Viterbi Alignments

The original log-likelihood that they would like to maximize requires summing over all of the possible alignments, which is unfortunatelly intractable (Valiant, 1979). But if the alignment $a$ is a fixed number, then optimization becames much more feasible. Burkett and Klein (2008) finally presumed an optimal $a$ over the tree pair and then continued to find the maximum weight $w$ using an EM-like algorithm.

### 3.2 Pruning

Burkett and Klein (2008) used $k$-best lists during training and testing for both of their source and target tree sets, $T$ and $T'$. In order to reduce the search space over the whole tree set in the training set, they pruned the tree sets by the $F_1$ score, so from the best until the $k$th best tree will be kept in the training set. A same strategy was used on the test set as well, though there they rank the tree set by a different metric.

Later in the experiments they have showed that the best $k$ value for the training set was 25, and the best $k$ value for the test set was 500. Also the performance of the parser was limited more by the model's reliance on the baseline parser, rather by the errors from a small $k$ value.

## 4 A More Practical Application

The other paper from Huang et al. (2009) showed a potentially more practical application of the above discovery. Their strategy is that rather than parsing the whole bitext all the time, we only need to consult the other language when we have encoutered an ambiguity in the source language.

Huang et al. (2009) believed that the errors in an arc-standard parser are mainly caused by the conflicts between `SHIFT` and `REDUCE`(which is a different term and includes both the `LEFT-ARC` and the `RIGHT-ARC` action).

### 4.1 Bilingual Contiguity Features

Their application is based on these two following observation.

**Pro-Reduce**: For the top two stack words $s_{t-1}$ and $s_t$, their correct spans in the target language should also be contiguous if the preferred action is `REDUCE`. This is captured by the feature $c(s_{t-1}, s_t)$.

**Pro-Shift**: For the stack top word $s_t$ and the current word in the buffer $w_i$, their correct span in the target language should start from $S_t$ without necessarily ending at $w_i$. This is captured by the feature $c_R(s_t, w_i)$.

Hence, $c(s_{t-1}, s_t) = +$ and $c_R(s_t, w_i) = -$ suggest `REDUCE` should be take while $c(s_{t-1}, s_t) = -$ and $c_R(s_t, w_i) = +$ suggest the opposite. They have even designed a more discriminatory feature by combining these two sub features, $c(s_{t-1}, s_t) \circ c_R(s_t, w_i)$.

Huang et al. (2009) showed that these features can indeed capture conflicts between `SHIFT` and `REDUCE`, especially in the case when `SHIFT` was mis-executed by the parser.

## 5 Results

Burkett and Klein (2008) conducted their tests on the Penn Chinese Treebank and their bitext parser outperformed the state-of-the-art(2008) monolingual parser baseline by 2.5 $F_1$ at predicting English side trees and 1.8 $F_1$ score at predicting Chinese side trees. Plus, their performance on the Chinese treebank was the highest published numbers on the same corpora in 2008. Even in sentences that lack translations, their parser can still get higher $F_1$ scores.

Following their encouraging results, the parser made by Huang et al. (2009) raised the $F_1$ score from their baseline by 0.6 on both the English and the Chinese corpora with negligible efficiency overhead (6%). However their result on the Chinese side trees didn't outperformed the Berkeley parser, which they believe was caused by the fact that they had engineered their features on English data instead of Chinese data.

## 6 Conclusion

Both papers demostrated techniques and potentials in bitext parsing, especially for languages that are largely different from each other. Their results were positive and encouraging as both of their theories are proved to outperform their baselines.

In addition, Burkett and Klein (2008) had discussed that how could Machine Translation could benefit from bitext parsing aswell, while Huang et al. (2009) did a great analysis on typical conflicts for an arc-standard parser and showed that an arc-standard parser might be as good as an arc-eager one. Following on their contribution, more future works should be promising in this field.

## References


Burkett, D. and Klein, D. (2008). Two languages are better than one (for syntactic parsing). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 877–886, Stroudsburg, PA, USA. Association for Computational Linguistics.

Huang, L., Jiang, W., and Liu, Q. (2009). Bilingually-constrained (monolingual) shift-reduce parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1222–1231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Valiant, L. G. (1979). The complexity of computing the permanent. *Theor. Comput. Sci.*, 8:189–201.

Xue, N., Chiou, F.-D., and Palmer, M. (2002). Building a large-scale annotated chinese corpus. In *COLING 2002: The 19th International Conference on Computational Linguistics*.