Iterative Alignment of Syntactic Structures for a Bilingual Corpus

Ralph Grishman

Computer Science Department
New York University

Abstract

Alignment of parallel bilingual corpora at the level of syntactic structure holds the promise of being able to discover detailed bilingual structural correspondences automatically. This paper describes a procedure for the alignment of regularized syntactic structures, proceeding bottom-up through the trees. It makes use of information about possible lexical correspondences, from a bilingual dictionary, to generate initial candidate alignments. We consider in particular how much dictionary coverage is needed for the alignment process, and how the alignment can be iteratively improved by having an initial alignment generate additional lexical correspondences for the dictionary, and then using this augmented dictionary for subsequent alignment passes.

Introduction

The process of aligning bilingual corpora can provide valuable information about the source and target languages and about bilingual correspondences. This alignment can be done at several levels. There is already a considerable literature on performing sentence-level alignment and identifying word-level correspondences (for example, [Church 93], [Chen 93], and works cited therein).

Our own work starts with a corpus which has been aligned at the sentence level, and considers the problem of alignment at the level of regularized syntactic structure — a level corresponding approximately to "deep structure" or the F-structure of lexical-functional grammar. Previous studies have shown that at this level, which abstracts

away some of the most apparent surface differences between languages, there is a considerable parallel between language structures [Harris 68, Teller et al. 88].

Alignment at this level serves several purposes. It can be used to identify vocabulary correspondences in a more focused way than sentence-level alignment (thus permitting, for example, identification of lexical correspondences from a single example). It can be used to disambiguate syntactic analyses in one language, using information from the corresponding sentence in the other language [Utsuro et al. 92, Matsumoto et al. 93]. And it can be used to identify correspondences at the level of syntactic case frames and larger syntactic structures, as would be required for a transfer-based machine translation system [Kaji et al. 92, Grishman and Kosaka 92]. The latter has been our principal motivation in developing this alignment procedure.

In the next section, we consider our motivation in somewhat more detail, focusing on the selection of the appropriate level of analysis at which to perform the alignment. The sections which follow describe the alignment algorithm itself, and some of the evaluations which we have made of the algorithm.

Level of Analysis

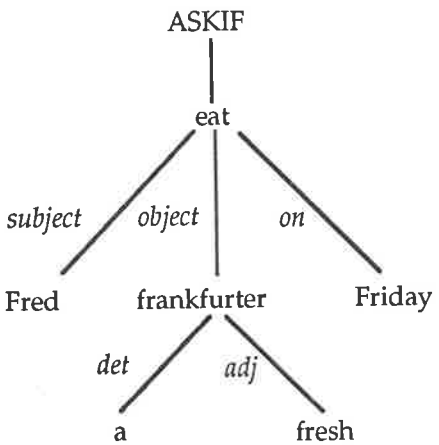
In developing language analysis systems we are always faced with the problem of having to encode large amounts of information. In analyzing text in a single language, for example, we are faced with the problem of capturing selectional constraints — information on the meaningful or allowable combinations of words. In machine translation, using a transfer-based approach, we are faced with the need to specify a large number of rules to map source language into target language structures.

These hurdles are now beginning to be overcome through the use of corpus-based discovery techniques. For monolingual analysis, there have now been several successful efforts at extracting selectional patterns from corpora [Sekine et al. 92, Chang et al. 92, Grishman and Sterling 92]. In the realm of machine translation, there have been two avenues of development. On the one hand, the work on Example-Based Machine Translation [Sato and Nagao 90, Sumita and Iida 91] and on Bilingual Knowledge Bases [Sadler and Vendelman 90] has shown that collections of manually-selected, syntactically analyzed bilingual examples can be an effective source of translation information, substituting for explicitly prepared transfer rules. On the other hand, the

work on Statistically-Based Machine Translation [Brown et al. 90] has shown that bilingual correspondences extracted automatically from corpora, with minimal syntactic processing, can be an effective base for a translation system. Our goal is to combine these two approaches by automatically extracting *structural* correspondences through the alignment of syntactically analyzed and regularized corpora.

Alignment at the word level, as was originally done by IBM, poses difficulties because the correspondences across languages can be quite complex. Using surface syntactic structures, as in some of the EBMT work and the recent work at Hitachi [Kaji et al. 92], simplifies the correspondences and hence the task of alignment. A regularization of the syntactic structures, for example to introduce a single representation of different clausal structures, further improves the correspondence across languages and thus the potential for a discovery procedure to automatically acquire these correspondences from a limited training sample.

In choosing a level of representation, we have sought to perform whatever regularizations can be stated in terms of general syntactic categories. Thus, in English, declarative and interrogative forms, active and passive clauses, relative and reduced relative clauses, are all reduced to a single form. In the resulting form (like f-structure), the basic structure consists of a head and a set of operands in particular syntactic roles, such as *subject, object, indirect object,* etc. for clauses, and *determiner, numeric quantifier, adjectival modifier*, etc. for noun phrases. For example, "Did Fred eat a fresh frankfurter on Friday?" would be represented as



Alignment Algorithm

For our procedure, we assume that the texts have already been aligned at the sentence level, and that both the source and target language texts have been syntactically analyzed and regularized. For the experiments reported here, no additional (selectional) constraints have been applied during parsing, so each source and target sentence will typically have a plurality of parses.

We also assume that we have available a bilingual dictionary which lists typical translations for many of the words in the corpus. We shall consider later just how many are required, but we do not require that the dictionary include translations for all the words, or only the translations used in the corpus. This information might be extracted from a commercial bilingual dictionary, or could itself be derived from a sentence-aligned corpus in an initial stage of processing. We may also have available information about correspondences between role names in the source and target trees.

Given a source and a target sentence tree, an *alignment* is a pairing of a subset of the nodes in the source tree with a subset of the nodes in the target tree. To represent the alignment, we number the nodes in the source tree, $1,...,N_S$, and the nodes in the target tree, $1,...,N_T$; an alignment is then a set of pairs $<s_i,t_i>$, $i=1,...,N_A$, indicating that node s_i of the source tree has been paired with node t_i of the target tree. For an alignment to be well-formed, we require that the relation of dominance in the tree be preserved in the mapping from source nodes to corresponding target nodes; that is, if the alignment includes $<s_i,t_i>$ and $<s_j,t_j>$ and s_i dominates s_j in the source tree, then t_i must dominate t_j in the target tree. (This condition is imposed so that, once correspondences have been identified, the trees can be chopped up into corresponding source and target subtrees.)

The minimal criterion for establishing a correspondence between nodes s_i and t_i is that either

- t_i is a possible translation of s_i as recorded in the bilingual dictionary
- there are one or more pairs $<s_j, t_j>$ in the alignment such that s_i dominates s_j and t_i dominates t_j

- there is a pair in the alignment, $<s_j, t_j>$ such that s_j immediately dominates s_i, t_j immediately dominates t_i, and the role filled by t_i is a possible translation of the role filled by s_i

These minimal criteria would allow for a large number of alternative alignments, so we assign a score to each alignment and select the highest scoring alignment. The score of an alignment is the sum of the scores of the individual correspondences making up the alignment. The score of an individual alignment $<s_i, t_i>$ is based in turn on four terms:

- whether t_i is a possible translation of s_i
- whether s_i dominates any other nodes in the alignment
- the distance from s_i to the other nodes in the alignment which are dominated by s_i (this has a negative weight: nodes which immediately dominate other corresponding nodes are preferred)
- for each node t_j in the alignment which is immediately dominated by t_i, whether the role filled by t_j is a possible translation of the role filled by the corresponding node s_j

The search for alignments proceeds bottom up through the source tree: for each source node, the procedure identifies possible corresponding target nodes, and generates an alignment, or extends previously hypothesized alignments, using each possible correspondence. A form of beam search is employed: a score is associated with each alignment, and only alignments whose score is within some beam width Δ of the score of the best alignment are retained.

When there are multiple parses of the source and target sentence, the alignment procedure is applied between each source parse and each target parse, and selects the source parse and the target parse which together yield the highest-scoring alignment. Unless there are parallel syntactic ambiguities in the source and target sentence, this process can be used to disambiguate (or at least reduce the ambiguity in) the source and target sentences.

Evaluation

For our initial evaluation of this alignment algorithm, we have selected some relatively simple texts: three chapters (73 sentences) from an introductory Spanish textbook, *El Camino Real* [Jarrett and McManus 58], along with English translations of these chapters. One of the byproducts of the alignment process is the selection of a preferred (best-aligning) source language parse, and we have used this as our initial evaluation measure. This is nearly the same measure which has been used in [Matsumoto et al. 93] for the evaluation of their alignment algorithm.

Table 1 shows the improvement in parse accuracy by using the alignment procedure. Without the procedure, the first parse is correct for 43% of the sentences; using the alignment procedure to select a parse yields a correct parse 59% of the time (Table 1, last row).

Method of selecting parse	Percentage of Correct Parses
No alignment	43%
Alignment, using 1/8 of textbook	48%
Alignment, using 1/3 of textbook	52%
Alignment, using entire textbook	59%

Table 1. Parse quality as a function of dictionary size for alignment algorithm.

This first experiment used as a bilingual dictionary the entire dictionary provided with the textbook. To gauge the extent to which successful alignment depended on adequate dictionary coverage, we repeated the alignment procedure using truncated dictionaries, first with 1/3 of the full dictionary, then with 1/8 of the full dictionary. As Table 1 shows, the quality of the alignments correlated with the size of the dictionary.

These experiments indicated the importance of having a procedure which is robust with respect to gaps in the bilingual dictionary. Even the dictionary provided with the textbook did not provide complete coverage, and considerably larger gaps could be

expected when the experiment is extended to use a broad-coverage bilingual dictionary and more complex texts. We therefore implemented an *iterative* alignment algorithm. During one pass through the texts, the procedure collects the correspondences from the best alignment of each sentence. At the end of the pass, it extracts the word correspondences which did not appear in the bilingual dictionary, and adds them to the bilingual dictionary. It also extracts the role correspondences and adds them, along with frequency information, to the table of role correspondences. This extended dictionary and table of role correspondences is then used in the next pass in aligning the text. (Analogous iterative algorithms have been described for *sentence* alignment, in which an initial alignment is used to estimate lexical correspondence probabilities, and these are then used to obtain an improved alignment [Chen 93]).

Through a series of such iterations, the coverage of the bilingual dictionary and table of role correspondences is gradually increased until a limiting state is reached. This is reflected in gradually improving scores on the parsing metric, as shown in Table 2. We began by using only one-eighth of the original dictionary. By the third iteration, the alignments are as good as those obtained with the full original dictionary (no further improvements were obtained by additional iterations).

Iteration Number	Percentage of Correct Parses
1	48%
2	53%
3	59%

Table 2. Improvement of parse quality through iterative alignment.

Discussion

A comparison of our methods with those adopted at Hitachi [Kaji et al. 92] and those adopted at Kyoto and Nara [Utsuro et al. 92, Matsumoto et al. 93] is instructive in understanding some of the alternatives possible in structural alignment.

We noted one difference earlier: the alignment at Hitachi is based on surface structure, whereas our work, and the work at Kyoto and Nara, involves the alignment of "deeper", functional syntactic structures.

There are differences in what constitutes an alignment. Our notion of alignment is consistent with that presented formally in [Matsumoto et al. 93]. For both groups, an alignment is a relation between complete source and target language trees, which respects the dominance relation in the tree (if nodes s_1 and t_1 correspond in the alignment, and so do s_2 and t_2, and s_1 dominates s_2, then t_1 must dominate t_2). In contrast, in Hitachi's approach the alignment of each source tree node to a target tree node is considered independently, and is not directly affected by the alignment of other nodes. (A choice of node alignments, however, may resolve ambiguous word alignments, and therefore indirectly affect subsequent node alignments; as a result, one would expect that in most cases the individual node alignments could be integrated into a tree alignment.)

These differences reflect different goals for the alignment process. The work at Kyoto and Nara has focused on the resolution of syntactic ambiguity. The work at NYU seeks to identify individual structural correspondences within the analysis trees. Both therefore require alignments between tree structures. The Hitachi group, in contrast, builds transfer patterns involving word sequences with limited phrase-structure annotation; these can be constructed by identifying individual correspondences, without aligning entire tree structures.

There are also marked differences in the procedures used to produce the alignments. In the work at Kyoto and Nara, the alignments are built top-down, using a branch-and-bound (backtracking) algorithm to find the best match. The alignment procedure at Hitachi, in contrast, operates bottom-up; it starts by identifying possible word correspondences and then aligns phrases (nodes) of gradually increasing length. It appears that decisions regarding node alignment are made deterministically. This approach fits well with the notion of treating the node alignments independently.

We have chosen to use an alignment algorithm which, like Hitachi's, operates primarily bottom-up. This decision was motivated in part by our earlier studies of parallel bilingual programming language manuals, which indicated that syntactic tree

correspondences were usually very close at the bottom of the tree, for the most sublanguage-specific material, while the trees could diverge considerably at the top (where general vocabulary such as "We will see that ..." was used). Matsumoto et al. note that their procedure encounters some difficulty if the roots of the source and target tree are quite different. In addition, the bottom-up algorithm should be able to handle quite naturally situations where a single source sentence corresponds to multiple target language sentences.

Our choice of a bottom-up algorithm was also motivated in part by considerations of efficiency. The top-down branch-and-bound algorithm can find the optimal match, but because the search space of possible matches is so large, it may take a very long time to do so. Our bottom-up match, guided by the word correspondences and employing limited backtracking (beam search), is not guaranteed to find an optimal alignment, but it appears that it can find acceptable alignments with more limited search. There are, however, cases where the pure bottom-up strategy behaves poorly. This shortcoming is particularly evident in sentences with multiple conjunctions, where a number of low-level alignments will be constructed, most of which will be discarded (due to low scores) when the top levels of the tree are reached (our training texts, while generally syntactically fairly simple, make heavy use of conjunction, presumably because it would be easy for beginning language learners to understand.)

To improve efficiency, we are now experimenting with a combination of top-down and bottom-up search. We begin by proceeding top-down, starting from the root, and continuing so long as there is a close lexical and structural match between source and target trees. When the top-down match stops (because there is some divergence between source and target trees), the remainder of the trees with be matched bottom-up using the procedure previously described.

Application: Transfer Rule Discovery Procedures

As we noted earlier, our objective in creating these alignments is to automatically extract transfer rules from the bilingual corpus. Once an alignment has been created, the source and target trees are "cut" at the nodes in the alignment, producing a set of source tree fragments and target tree fragments. If every node is in the alignment, each tree fragment will be a single level of the tree, indicating how a head plus a set of

syntactic roles in the source language is mapped into a head plus roles in the target language. If the alignment does not include every node, the mapping may go from a single level in the source language to two or more levels in the target language, or vice versa — a "structural transfer". These corresponding tree fragments are then collected and generalized to form the transfer rules of a translation system.

We have completed a rudimentary system of this form for producing translations from Spanish to English. However, because of the simplicity of the sentences in our current training corpus (the first few chapters of our Spanish textbook), almost no structural transfer is needed (once the text is parsed, translation is nearly direct), and so the capability of this approach to acquire and generalize such structural rules is not yet seriously tested. We intend in the near future to extend our training corpus to larger portion of this textbook and to other texts in order to properly gauge the power of our procedure in acquiring structural transfer rules.

Acknowledgment

The work reported in this paper was supported by the National Science Foundation under Grant IRI-9303013.

References

[Brown et al. 90] P. F. Brown, J. Cocke, S. A. DellaPietra, V. J. DellaPietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roosin. A statistical approach to machine translation. *Computational Linguistics* **16** (2), 1990.

[Chang et al. 92] J.-S. Chang, Y.-F. Luo, and K.-Y. Su. GPSM: a generalized probabilistic semantic model for ambiguity resolution. *Proc. 30th Annl. Meeting Assn. for Computational Linguistics*, Newark, DE, 1992, 177-184.

[Chen 93] S. Chen, Aligning sentences in bilingual corpora using lexical information. *Proc. 31st Annl. Meeting Assn. Computational Linguistics,*, Columbus, Ohio, June 1993, 9-16.

[Church 93] K. W. Church, Char_align: a program for aligning parallel texts at the character level. *Proc. 31st Annl. Meeting Assn. Computational Linguistics,*, Columbus, Ohio, June 1993, 1-8.

[Grishman and Kosaka 92] R. Grishman and M. Kosaka. Comparing rationalist and empiricist approaches to machine translation. *Proc. Fourth Int'l Conf. on Theoretical and Methodological Issues in Machine Translation*, Montreal, June 1992.

[Grishman and Sterling 92] R. Grishman and J. Sterling. Acquisition of selectional patterns. *Proc. 14th Int'l Conf. on Computational Linguistics,* Nantes, France, 1992.

[Jarrett and McManus 58] E. M. Jarrett and B. J. M. McManus. *El Camino Real, Book One*. Boston: Houghton Mifflin, 1958.

[Harris 68] Z. Harris, *Mathematical Structures of Language*. New York: Wiley Interscience, 1968.

[Kaji et al. 92] H. Kaji, Y. Kida, and Y. Morimoto. Learning translation templates from bilingual text. *Proc. 14th Int'l Conf. on Computational Linguistics*, Nantes, 1992, 672-678.

[Matsumoto et al. 93] Y. Matsumoto, H. Ishimoto, T. Utsuro, and M. Nagao. Structural matching of parallel texts. *Proc. 31st Annl. Meeting Assn. Computational Linguistics,*, Columbus, Ohio, June 1993, 23-30.

[Sadler and Vendelman 90] V. Sadler and R. Vendelman. Pilot implementation of a bilingual knowledge bank. *Proc. 13th Int'l Conf. on Computational Linguistics*, Helsinki, Finland, 1992, 449-451.

[Sato and Nagao 90] S. Sato and M. Nagao. Toward memory-based translation. *Proc. 13th Int'l Conf. Computational Linguistics,* Helsinki, Finland, 1990, 247-252.

[Sekine et al. 92] S. Sekine, J. Carroll, A. Ananiadou, and J. Tsujii. Automatic learning for semantic collocation. *Proc. Third Conf. Applied Natural Language Processing*, Trento, Italy, 1992, 104-110.

[Sumita and Iida 91] E. Sumita and H. Iida. Experiments and prospects of example-based machine translation. *Proc. 29th Annl. Meeting Assn. for Computational Linguistics*, Berkeley, Ca, 1991.

[Teller et al. 88] V. Teller, M. Kosaka, and R. Grishman. A comparative study of Japanese and English sublanguage patterns. *Proc. Second Int'l Conf. on Theoretical and Methodological Issues in Machine Translation*, Pittsburgh, PA, 1988.

[Utsuro et al. 92] T. Utsuro, Y. Matsumoto, and M. Nagao. Lexical knowledge acquisition from bilingual corpora. *Proc. 14th Int'l Conf. on Computational Linguistics*, Nantes, 1992, 581-587.