

Exploiting Parallel Corpora for Monolingual Grammar Induction —A Pilot Study

Jonas Kuhn

The University of Texas at Austin
Department of Linguistics
Austin, TX 78712, USA
jonask@mail.utexas.edu

Abstract

This paper presents results from a pilot study on ways of exploiting statistical word alignment for grammar induction. Following a scheme proposed in (Kuhn, 2004), we use GIZA++-word alignment from the multiple parallel texts in the Europarl corpus for the identification of string spans that cannot be constituents in one of the languages. This information is exploited in monolingual PCFG grammar induction for that language. Besides the aligned corpus, no other resources are required.

1. Introduction

There have been a number of recent studies exploiting parallel corpora in bootstrapping of monolingual analysis tools. In the “information projection” approach (e.g., (Yarowsky and Ngai, 2001)), statistical word alignment is applied to a parallel corpus of English and some other language F for which no tagger/morphological analyzer/chunker etc. (henceforth simply: analysis tool) exists. A high-quality analysis tool is applied to the English text, and the statistical word alignment is used to project a (noisy) target annotation to the F version of the text. Robust learning techniques are then applied to bootstrap an analysis tool for F , using the annotations projected with high confidence as the initial training data. (Confidence of both the English analysis tool and the statistical word alignment is taken into account.) The results that have been achieved by this method are very encouraging.

Will the information projection approach also work for less shallow analysis tools, in particular full syntactic parsers? An obvious issue is that one does not expect the phrase structure representation of English (as produced by state-of-the-art treebank parsers) to carry over to less configurational languages. Therefore, (Hwa et al., 2002) extract a more language-independent dependency structure from the English parse as the basis for projection to Chinese. From the resulting (noisy) dependency treebank, a dependency parser is trained using the techniques of (Collins, 1999). (Hwa et al., 2002) report that the noise in the projected treebank is still a major challenge, suggesting that a future research focus should be on the filtering of (parts of) unreliable trees and statistical word alignment models sensitive to the syntactic projection framework.

Our hypothesis is that the quality of the resulting parser/grammar for language F can be significantly improved if the training method for the parser is changed to accommodate for training data which are in part unreliable. The experiments we report in this paper focus on a specific part of the problem: we replace standard treebank training with an Expectation-Maximization (EM) algorithm for PCFGs, augmented by weighting factors for the reliability of training data, following the approach of (Nigam et al., 2000), who apply it for EM training of a text classifier. The

factors are only sensitive to the constituent/distituent status of each span of the string in F (cp. (Klein and Manning, 2002)). The constituent/distituent status is derived from an aligned parallel corpus using the scheme of (Kuhn, 2004) (compare section 2.). We use the Europarl corpus (Koehn, 2002), and the statistical word alignment was performed with the GIZA++ toolkit (Al-Onaizan et al., 1999; Och and Ney, 2003).¹

For the current experiments we assume no pre-existing parser for any of the languages, contrary to the information projection scenario. While better absolute results could be expected using one or more parsers for the languages involved, we think that it is highly informative to run a pilot study that isolates the effect of using crosslinguistic word order divergences as prior knowledge about the constituent structure of a language. This prior knowledge is exploited in an EM learning approach (section 3.). Not using a parser for some languages also makes it possible to compare various language pairs at the same level, and since we don't need English as the most reliable basis of projection, we can in particular run grammar induction experiments for English (section 4.), which facilitates evaluation against a treebank (section 5.).

2. Cross-language order divergences

The English-French example in figure 1 gives a simple illustration of the partial information about constituency that a word-aligned parallel corpus may provide. The en bloc reversal of subsequences of words provides strong evidence that, for instance, [*moment the voting*] or [*aura lieu à ce*] do not form constituents.

At first sight it appears as if there is also clear evidence for [*at that moment*] forming a constituent, since it fully covers a substring that appears in a different position in French. Similarly for [*Le vote aura lieu*]. However, from the distribution of contiguous substrings alone we cannot distinguish between the two types of situations sketched in (1) and (2): a string that is contiguous under projection, like e_1e_2 (1) may be a true constituent, but it may also be a non-constituent part of a larger constituent as in L_1 in (2).

¹The software is available at
<http://www.isi.edu/~och/GIZA++.html>

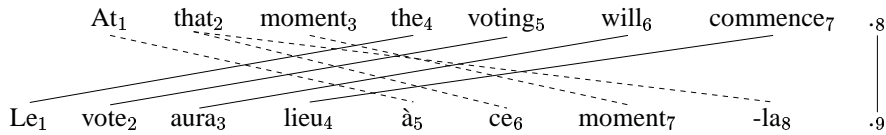
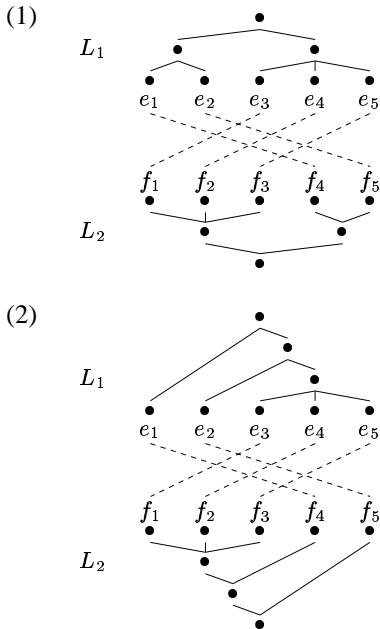


Figure 1: Alignment example



(Kuhn, 2004) provides a detailed discussion on the formal conditions for hypothesizing reliable non-constituent spans in a word-aligned corpus.

The core idea is to mark the boundary between contiguous word blocks (e.g., between e_2 and e_3 (1) or (2)). Then, spans of words crossing such boundaries without exhaustively covering one of the adjacent blocks are excluded from constituent status, i.e., we mark them as *distituents*.

Mild divergences are best. As should be clear, our scheme for detecting clues for non-constituency (i.e., information about distituents) relies on the occurrence of reorderings of constituents in translation. If two languages have the exact same structure (and no paraphrases whatsoever are used in translation), the approach does not gain any information from a parallel text. However, this situation does not occur realistically. If on the other hand, massive reordering occurs without preserving *any* contiguous sub-blocks, the approach cannot gain information either. The ideal situation is in the middleground, with a number of mid-sized blocks in most sentences.

3. EM grammar induction with weighting factors

The distituent identification scheme introduced in (Kuhn, 2004) and reviewed briefly in the previous section can be used to hypothesize a fairly reliable exclusion of constituency for many spans of strings from a parallel corpus. Besides a statistical word alignment, no further resources are required.

In order to make use of this scattered (non-)constituency information in grammar induction, a semi-supervised approach is needed that can fill in the (potentially large) areas for which no prior information is available. For the present experiments we decided to choose a conceptually simple such approach, with which we can build on substantial existing work in grammar induction: we construe the learning problem as PCFG induction, using the inside-outside algorithm, with the addition of weighting factors based on the (non-)constituency information. This use of weighting factors in EM learning follows the approach discussed in (Nigam et al., 2000).

For our pilot study, the conceptual simplicity and the availability of efficient implemented open-source systems of a PCFG induction approach outweighs the disadvantage of potentially poorer overall performance than one might expect from some other approaches.

The PCFG topology we use is a binary, entirely unrestricted X-bar-style grammar based on the Penn Treebank POS-tagset (expanded as in the TreeTagger by (Schmid, 1994)). All possible combinations of projections of POS-categories X and Y are included following the schemata in (3). This gives rise to 13,110 rules.

- (3) a. $XP \rightarrow X$
- b. $XP \rightarrow XP YP$
- c. $XP \rightarrow YP XP$
- d. $XP \rightarrow YP X$
- e. $XP \rightarrow X YP$

We tagged the English version of our training section from the Europarl corpus with the TreeTagger and used the strings of POS-tags as the training corpus for the inside-outside algorithm.²

We based our EM training algorithm on Mark Johnson's implementation of the inside-outside algorithm.³ The initial parameters on the PCFG rules are set to be uniform. In the iterative induction process of parameter reestimation, the current rule parameters are used to compute the expectations of how often each rule occurred in the parses of the training corpus, and these expectations are used to adjust the rule parameters, so that the likelihood of the training data is increased. When the probability of a given rule drops below a certain threshold, the rule is excluded from the grammar. The iteration is continued until the increase in likelihood of the training corpus is very small.

²Note that it is straightforward to apply our approach to a language for which no taggers are available if an unsupervised word clustering technique is applied first.

³<http://cog.brown.edu/~mj/>

Weight factors. The inside-outside algorithm is a dynamic programming algorithm that uses a chart in order to compute the rule expectations for each sentence. We use the information obtained from the parallel corpus as discussed in section 2. (and more extensively in (Kuhn, 2004)) as prior information (in a Bayesian framework) to adjust the expectations that the inside-outside algorithm determines based on its current rule parameters. Note that this prior information is information about string spans of (non-)constituents – it does not tell us anything about the categories of the potential constituents affected. It is combined with the PCFG expectations as the chart is constructed. For each span in the chart, we get a weight factor that is multiplied with the parameter-based expectations. In the simplest model, we use the factor 0 for spans that are clear distituent, and factor 1 for all other spans; in other words, parses involving a distituent are cancelled out. We also used versions of the weight factors in which a number of levels is applied: distituent is assigned factor 0.01, likely distituent factor 0.1, neutral spans 1, and likely constituents factor 2.⁴ The multi-level factor system turns out to outperform the simple distituent scheme.

4. Experiments

We applied GIZA++ (Al-Onaizan et al., 1999; Och and Ney, 2003) to word-align parts of the Europarl corpus (Koehn, 2002) for English and all other 10 languages. For the experiments we report in this paper, we only used the 1999 debates, with the language pairs of English combined with Finnish, French, German, Greek, Italian, Spanish, and Swedish.

For computing the weight factors we used a two-step process implemented in Perl, which first determines the location of boundaries between contiguous word blocks under cross-language word alignment. (5) shows the internal representation of the block structure for (4). L and R are used for the beginning and end of blocks, where it is unambiguous because there are no adjacent zero-fertility words (i.e., words for which the word alignment does not specify a correspondent). The notation l and r is used where zero-fertility word make the representation ambiguous. Words whose correspondents are in the same word order sequence are encoded as *, zero fertility words as -; A and B are used for the first block in a sentence instead of L and R, unless it arises from “relocation”, which increases likelihood for constituent status (likewise for the last block: Y and Z).

(4) la parole est à m. graefe zu baringdorf
pour motiver la demande
NULL ({ 3 4 11 }) mr ({ 5 }) graefe ({ 6
}) zu ({ 7 }) baringdorf ({ 8 }) has ({
}) the ({ 1 }) floor ({ 2 }) to ({ 9 })
explain ({ 10 }) this ({ }) request ({
12 })

(5) [L**r-lRY*-*Z]

⁴The factor weights were chosen empirically; but it can be expected that in the future, a more systematic technique using a set of held-out data will lead to further improvements.

The second step for computing the weight factors creates a chart of all string spans over the given sentence and marks for each span whether it is a distituent, possible constituent or likely distituent, based on the location of boundary symbols. (For instance *zu Baringdorf has the* is marked as a distituent; *the floor* and *has the floor* are marked as likely constituents.) The tests are implemented as simple regular expressions. The chart of weight factors is represented as an array which is stored in the training corpus file along with the sentences. We combine the weight factors from various languages, since each of them may contribute distinct (non-)constituent information. The inside-outside algorithm reads in the weight factor array and uses it in the computation of expected rule counts.

We used the probability of the statistical word alignment as a confidence measure to filter out unreliable training sentences. Due to the conservative nature of the constituent/distituent information we extract from the alignment, the results indicate however that filtering is not necessary.

5. Evaluation

For evaluation, we used the PCFG resulting from the training described in section 4. in order to find the best parse for each test sentence according to the model. For this, we ran the trained grammar with the Viterbi algorithm⁵ on parts of the Wall Street Journal (WSJ) section of the Penn Treebank and compared the predicted tree structure with the gold standard treebank annotation. The evaluation criteria we apply are unlabeled bracketing precision and recall (and crossing brackets). We follow an evaluation criterion that (Klein and Manning, 2002, footnote 3) discuss for the evaluation of a not fully supervised grammar induction approach based on a binary grammar topology: bracket multiplicity (i.e., non-branching projections) is collapsed into a single set of brackets (since what is relevant is the constituent structure that was induced).⁶ For comparison, we provide baseline results that a uniform left-branching structure and a uniform right-branching structure (which encodes some non-trivial information about English syntax) would give rise to. As an upper boundary for the performance that a binary grammar can achieve on the WSJ, we present the scores for a minimal binarized extension of the gold-standard annotation.

The results we can report at this point are based on a comparatively small training set.⁷ So, it may be too early for conclusive results. (An issue that arises with the small training set is that smoothing techniques would be required to avoid overtraining, but these tend to dominate the test application, so the effect of the parallel-corpus based information cannot be seen so clearly.) But we think that the

⁵We used the LoPar parser (Schmid, 2000) for this.

⁶Note that we removed null elements from the WSJ, but we left punctuation in place. We used the EVALB program for obtaining the measures, however we preprocessed the bracketings to reflect the criteria we discuss here.

⁷This is not due to scalability issues of the system; we expect to be able to run experiments on rather large training sets. Since no manual annotation is required, the available resources are practically indefinite.

System	Unlab. Prec.	Unlab. Recall	F ₁ -Score	Crossing Brack.
Left-branching	30.4	35.8	32.9	3.06
Right-branching	36.2	42.6	39.2	2.48
Standard PCFG induction	42.4	64.9	51.3	2.2
PCFG trained with constituent/distituent weight factors from Europarl corpus	47.8	72.1	57.5	1.7
Upper limit	66.08	100.0	79.6	0.0

Figure 2: Scores for test sentences up to length 10.

results are rather encouraging.

As the table in figure 2 shows, the PCFG we induced based on the parallel-text derived weight factors reaches 57.5 as the F₁-score of unlabeled precision and recall. We show the scores for an experiment without smoothing, trained on about 3,000 sentences. Since no smoothing was applied, the resulting coverage (with low-probability rules removed) on the test set is about 80%. It took 74 iterations of the inside-outside algorithm to train the weight-factor-trained grammar; the final version has 1005 rules.

For comparison we induced another PCFG based on the same X-bar topology without using the weight factor mechanism. This grammar ended up with 1145 rules after 115 iterations. The F₁-score is only 51.3 (while the coverage is the same as for the weight-factor-trained grammar).

6. Discussion

This paper presented a pilot study on ways of using parallel corpora as the only resource in the creation of a monolingual analysis tools. We believe that in order to induce high-quality tools based on statistical word alignment, the training approach for the target language tool has to be able to exploit islands of reliable information in a stream of potentially rather noisy data. We experimented with an initial idea to address this task, which is conceptually simple and can be implemented building on existing technology: using the notion of word blocks projected by word alignment as an indication for (mainly) impossible string spans. Applying this information in order to impose weighting factors on the EM algorithm for PCFG induction gives us a first, simple instance of the “island-exploiting” system we think is needed. More sophisticated models may make use some of the experience gathered in these experiments.

The conservative way in which cross-linguistic relations between phrase structure is exploited has the advantage that we don’t have to make unwarranted assumptions about direct correspondences among the majority of constituent spans, or even direct correspondences of phrasal categories. The technique is particularly well-suited for the exploitation of parallel corpora involving multiple languages like the Europarl corpus. Note that nothing in our methodology made any language particular assumptions; future research has to show whether there are language pairs that are particularly effective, but in general the technique should be applicable for whatever parallel corpus is at hand.

A number of studies are related to the work we presented, most specifically work on parallel-text based “information projection” for parsing (Hwa et al., 2002), but also

grammar induction work based on constituent/distituent information (Klein and Manning, 2002) and (language-internal) alignment-based learning (van Zaanen, 2000). However to our knowledge the specific way of bringing these aspects together which we proposed in (Kuhn, 2004) is new.

7. References

- Al-Onaizan, Yaser, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky, 1999. Statistical machine translation. Final report, JHU Workshop.
- Collins, M., 1999. A statistical parser for Czech. In *Proceedings of ACL*.
- Hwa, R., P. Resnik, and A. Weinberg, 2002. Breaking the resource bottleneck for multilingual parsing. In *Proceedings of LREC*.
- Klein, D. and C. Manning, 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of ACL*.
- Koehn, P., 2002. Europarl: A multilingual corpus for evaluation of machine translation. Ms., University of Southern California.
- Kuhn, Jonas, 2004. Experiments in parallel-text based grammar induction. Ms., The University of Texas at Austin.
- Nigam, K., A. K. McCallum, S. Thrun, and T. Mitchell, 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134.
- Och, F. J. and H. Ney, 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.
- Schmid, Helmut, 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*. Manchester, UK.
- Schmid, Helmut, 2000. Lopar: Design and implementation. Arbeitspapiere des Sonderforschungsbereiches 340, No. 149, IMS Stuttgart.
- van Zaanen, M., 2000. ABL: Alignment-based learning. In *COLING 2000 - Proceedings of the 18th International Conference on Computational Linguistics*.
- Yarowsky, D. and G. Ngai, 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of NAACL*.