Unsupervised Grammar Induction by Distribution and Attachment

David J. Brooks
School of Computer Science
University of Birmingham
Birmingham, B15 2TT, UK
d.j.brooks@cs.bham.ac.uk

Abstract

Distributional approaches to grammar induction are typically inefficient, enumerating large numbers of candidate constituents. In this paper, we describe a simplified model of distributional analysis which uses heuristics to reduce the number of candidate constituents under consideration. We apply this model to a large corpus of over 400000 words of written English, and evaluate the results using EVALB. We show that the performance of this approach is limited, providing a detailed analysis of learned structure and a comparison with actual constituent-context distributions. This motivates a more structured approach, using a process of attachment to form constituents from their distributional components. Our findings suggest that distributional methods do not generalize enough to learn syntax effectively from raw text, but that attachment methods are more successful.

1 Introduction

Distributional approaches to grammar induction exploit the principle of substitutability: constituents of the same type may be exchanged with one another without affecting the syntax of the surrounding context. Reversing this notion, if we can identify "surrounding context" by observation, we can hypothesize that word sequences occurring in that context will be constituents of the same type. Thus, distributional methods can be used to segment text into constituents and classify the results. This work focuses on distributional learning from raw text.

Various models of distributional analysis have been used to induce syntactic structure, but most use probabilistic metrics to decide between candidate constituents. We show that the efficiency of these systems can be improved by exploiting some properties of probable constituents, but also that this reliance on probability is problematic for learning from text. As a consequence, we propose an extension to strict distributional learning that incorporates more information about constituent boundaries.

The remainder of this paper describes our experiences with a heuristic system for grammar induction. We begin with a discussion of previous distributional approaches to grammar induction in Section 2 and describe their implications in Section 3. We then introduce a heuristic distributional system in Section 4, which we analyze empirically against a treebank. Poor system performance leads us to examine actual constituent-context distributions (Section 5), the implications of which motivate a more structured extension to our learning system, which we describe and analyze in Section 6.

2 Previous approaches

Distributional methods analyze text by *alignment*, aiming to find equivalence classes covering substitutable units. We align common portions of texts termed *contexts*, leaving distinct contiguous word-sequences, termed *expressions*. An expression and its context form an *alignment pattern*, which is de-

Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X),
pages 117–124, New York City, June 2006. ©2006 Association for Computational Linguistics

fined as:

$$C_{left} \mid \textit{Expression} \mid C_{right} \qquad \text{(AP1)}$$

From this alignment pattern, we can extract context-free grammar rules:

$$NT \rightarrow Expression_1 \vee ... \vee Expression_n \qquad (1)$$

While the definition of expression is straightforward, the definition of context is problematic. We would like as much context as possible, but word-sequence contexts become less probable as their length increases, making learning harder. Therefore, simple models of context are preferred, although the precise definition varies between systems.

Distributional approaches to grammar induction fall into two categories, depending on their treatment of nested structure. The first category covers Expectation-Maximization (EM) systems. These systems propose constituents based on analysis of text, then select a *non-contradictory combination* of constituents for each sentence that maximizes a given metric, usually parsing probability. EM has the advantage that constituent probabilities are only compared when constituents compete, which removes the inherent bias towards shorter constituents, which tend to have higher probability. However, EM methods are more susceptible to data sparsity issues associated with raw text, because there is no generalization during constituent proposal.

Examples of EM learning systems are Context Distribution Clustering (CDC) (Clark, 2001) and Constituent-Context Model (CCM) (Klein, 2005, Chapter 5), which avoid the aforementioned data-sparsity issues by using a part-of-speech (POS) tagged corpus, rather than raw text. Alignment Based Learning (ABL) (van Zaanen, 2000) is the only EM system applied directly to raw text. ABL uses minimal String-Edit Distance between sentences to propose constituents, from which the most probable combination is chosen. However, ABL is relatively inefficient and has only been applied to small corpora.

The second category is that of incremental learning systems. An incremental system analyzes a corpus in a bottom-up fashion: each time a new constituent type is found, it is inserted into the corpus to provide data for later learning. This has the advantage of easing the data-sparsity issues described above because infrequent sequences are clustered into more frequent non-terminal symbols. However, in incremental systems, constituents are compared directly, which can lead to a bias towards shorter constituents.

The EMILE system (Adriaans, 1999) learns *shallow* languages in an incremental manner, and has been applied to natural language under the assumption that such languages are shallow. Shallowness is the property whereby, for any constituent type in a language, there exist well-supported minimal units of that type. EMILE aligns complete sentences only, attempting to isolate minimal units, which are then used to process longer sequences. This method is efficient because alignment is non-recursive. However, as a consequence, EMILE offers only a limited treatment of nested and recursive structures.

A more comprehensive approach to learning nested structure is found in the ADIOS system (Solan et al., 2003). ADIOS enumerates all patterns of a given length, under the condition that each sequence must have non-empty contexts and expressions. These patterns are ranked using an information gain metric, and the best pattern at each iteration is rewritten into the graph, before pattern scanning begins again. ADIOS learns context-sensitive equivalence classes, but does not induce grammars, and has not been formally evaluated against treebanks.

Grammar induction systems are evaluated using standard metrics for parser evaluation, and in particular, the EVALB algorithm[1]. The above systems have been evaluated with respect to the ATIS treebank. Compared with supervised parsers, these systems perform relatively poorly, with the strictly unsupervised EMILE and ABL systems recovering 16.8% and 35.6% of constituent structure respectively. The partially-supervised systems of CDC and CCM perform better, with the latter retrieving 47.6% of the constituent structure in ATIS. However, the strictly unsupervised systems of ABL, EMILE and ADIOS have not been evaluated on larger corpora, in part due to efficiency constraints.

[1] There are known issues with parser evaluation, although a discussion of these issues is outside the scope of this paper, and the reader is referred to (Klein, 2005, Chapter 2). We assume the standard evaluation for comparison with previous work.

3 Issues for distributional learning

There are many issues with distributional learning, especially when learning from raw text. First, previous systems hypothesize and select constituents according to the probability of their contexts: ABL, EMILE and CCM use the probability of proposed equivalence classes, or the equivalent context probability; ADIOS uses an information gain metric, again favouring probable contexts. However, when learning from raw text, this preference for hypotheses with more probable contexts means that open-class words will seldom be considered as contexts. In POS-based learners, it is possible to align open-class POS contexts. These contexts are demonstrably important despite low word probabilities, which suggests that selecting contexts on the basis of probability will be limited in success.

The second problem relates to word-senses. Alignment proceeds by matching orthographic types, but these types can have numerous associated syntactic senses. For example, ‘to’ plays two distinct roles: infinitive marker or preposition. If we align using the orthographic type, we will often misalign words, as seen in the following alignment:

I gave it	to	the man	in	the grey jacket
John agreed	to	see me	in	20 minutes

Here, we are (mis)aligning a prepositional ‘to’, with an infinitive marker. The result would be a correctly identified noun-phrase, ‘the man’, and an incorrect structure, contradicting both the verb-group ‘to see’ and the noun-phrase ‘me’. This problem does not affect POS-based learning systems, as POS tags are unambiguously assigned.

Finally, grammar induction systems are typically inefficient, which prohibits training over large corpora. Distributional analysis is an expensive procedure, and must be performed for large numbers of word sequences. Previous approaches have tended to enumerate all alignment patterns, of which the best are selected using probabilistic metrics. However, given the preference for probable alignments, there is considerable wasted computation here, and it is on this issue that we shall focus.

4 A heuristic approach to alignment

Rather than enumerating all word sequences in a corpus, we propose a heuristic for guiding distributional systems towards more favourable alignment patterns, in a system called *Directed Alignment*. In this system, we define context as the ordered pair of left- and right-context for a given constituent, $\langle C_{left} - C_{right} \rangle$, where C_{left} and C_{right} are single-units. The atomic units of this system are words, but learned constituents may also act as context-units.

The probability of a pattern depends primarily on its contexts, since they are common to all matching sequences. We can reduce the task of finding probable alignments to simply finding probable context-pairs. However, we can reduce this further: for a context-pair to be probable, its components must also be probable. Therefore, rather than enumerating all patterns in the corpus, we direct the alignment procedure towards patterns where C_{left} and C_{right} are probable.

The first stage of direction creates an index for the corpus, compiling a list of unit types, where units are initially words. From this list of types, the most probable 1% are selected as *context-units*. These context-units are the only types allowed to fill the roles C_{left} and C_{right} in alignment patterns.

Alignments are created directly from the context-unit index. For each context-unit token cu in the index, we locate cu in the corpus and create an alignment pattern, such that cu is the left context (C_{left}). Next, we scan the sequence of words following cu, extending the alignment pattern until another context-unit cu' is found, or a fixed length threshold is exceeded. If cu' is found, it fills the role of right context (C_{right}), and the completed alignment pattern is cached; otherwise, the pattern is disregarded.

Direction permits two forms of valid expressions in the context $\langle cu - cu' \rangle$:

1. $nc_1 \ldots nc_n$, where each nc_i is a non-context
2. $c_1 \ldots c_n$, where each c_i is a context-unit

The first of these forms allows us to examine non-nested alignments. The second allows us to analyze nested alignments only after inner constituents have been learned. These constraints reduce the number of constituents under consideration at any time to a manageable level. As a result, we can scan very large numbers of alignment patterns with relatively little overhead.

As an example, consider the following sequence, with context units underlined:

> put the whole egg , all the seasonings and vegetables into the bowl and process for 10 seconds until smoothly pured .

This would be broken into non-recursive expressions[2]:

> (put) the (whole egg) , all the (seasonings) and (vegetables) into the (bowl) and (process) for (10 seconds) until (smoothly pureed) .

These expressions will be replaced by non-terminal unit representing the class of expressions, such that each class contains all units across the corpus that occur in the same context:

> NT0 the NT1 , all the NT2 and NT3 into the NT2 and NT4 for NT5 until NT6 .

Following this generalization nested structures can be discovered using the same process.

This approach has some interesting parallels with chunking techniques, most notably that of function-word phrase identification (Smith and Witten, 1993). This similarity is enforced by disallowing nested structures. Unlike chunking systems, however, this work will also attempt to recover nested structures by means of incremental learning.

4.1 Selecting alignment patterns

The direction process extracts a set of candidate alignments, and from this set we select the best alignment to rewrite as an equivalence class. Previous approaches offer a number of metrics for ranking constituents, based around constituent or context probability (ABL and CCM), Mutual Information (CDC), and information gain (ADIOS). We have implemented several of these metrics, but our experiences suggest that context probability is the most successful.

The probability of an alignment is effectively the sum of all path probabilities through the alignment:

$$P(C_{left}, C_{right}) = \Sigma P(path_{left,right}) \qquad (2)$$

where each $path_{left,right}$ is a unique word sequence starting with $left$ and ending with $right$, under the constraints on expressions described above. There is an important practical issue here: probability sums such as that in Equation 2 do not decrease when expressions are replaced with equivalence classes. To alleviate this problem, we rewrite the units when updating the distribution, but discard paths that match the current alignment. This prevents looping while allowing the rewritten paths to contribute to nested structures.

4.2 Generalizing expression classes

The model outlined above is capable of learning strictly context-sensitive constituents. While this does allow for nested constituents, it is problematic for generalization. Consider the following equivalence classes, which are proposed relatively early in Directed Alignment:

the	NT1	of
the	NT2	in

Here, the non-terminals have been assigned on the basis of context-pairs: NT1 is defined by $\langle the - of \rangle$ and NT2 is defined by $\langle the - in \rangle$. These types are distinct, although intuitively they account for simple noun-phrases. If we then propose an alignment pattern with NT1 as C_{left}, it must be followed by 'of', which removes any possibility of generalizing 'of' and 'in'.

We alleviate this problem by generalizing equivalence classes, using a simple clustering algorithm. For each new alignment, we compare the set of expressions with all existing expression classes, ranking the comparisons by the degree of overlap with the current alignment. If this degree of overlap exceeds a fixed threshold, the type of the existing class is assumed; otherwise, a new class is created.

4.3 Experiments, results and analysis

To evaluate our algorithm, we follow the standard approach of comparing the output of our system with that of a treebank. We use the EVALB algorithm, originally designed for evaluating supervised parsing systems, with identical configuration to that of (van Zaanen, 2000). However, we apply our algorithms to a different corpus: the written sub-corpus of the International Corpus of English, Great Britain Component (henceforth ICE-GB), with punctuation removed. This consists of 438342 words, in 22815 sentences. We also include a baseline instantiation

[2] For clarity, we have shown all alignments for the given sentence simultaneously. However, the learning process is incremental, so each alignment would be proposed during a distinct learning iteration.

System	UP	UR	F_1	CB
FWB	30.0	11.0	16.0	0.36
DA	23.3	8.0	11.9	0.30
$DA_{cluster}$	23.6	8.1	12.0	0.30

Table 1: EVALB results after 500 iterations of Directed Alignment applied to ICE-GB, showing both context-sensitive (DA) and clustered ($DA_{cluster}$) alignment. The columns represent Unlabeled Precision, Unlabeled Recall, Unlabeled F-Score and the proportion of sentence with crossing brackets respectively.

of our algorithm, which chunks text into expressions between function words, which we refer to as Function-Word Bracketing (FWB).

Table 1 summarizes the EVALB scores for two 500-iteration runs of Directed Alignment over ICE-GB: DA is the standard context-sensitive version of the algorithm; $DA_{cluster}$ is the version with context clustering. FWB precision is relatively low, with only 30% of proposed structures appearing in the treebank. Recall is even lower, with only 11% of structure retrieved. This is unsurprising, as no nested constructions are considered.

In comparison, both versions of Directed Alignment perform significantly worse, with $DA_{cluster}$ being only fractionally better than standard DA. Experiments over more learning iterations suggest that the performance of DA converges on FWB, with few nested constituents discovered. Both variants of the system produce very poor performance, with very little nested structure recovered. While these results seem discouraging, it is worth investigating system performance further.

		Recall (%)		
Category	Frequency	FWB	DA	$DA_{cluster}$
NP	117776	11.81	10.83	10.79
CL	28641	0.50	1.21	1.14
VP	50280	20.88	9.58	9.89
PP	42134	0.10	0.67	0.73
SUBP	7474	1.10	11.05	11.15
NONCL	1919	4.27	22.98	22.98

Table 2: Constituent retrieval results for Function-Word Bracketing (FWB) and Directed Alignment (DA and $DA_{cluster}$), categorized by gold-type

(a) DA, top 5 noun-matches of 271

Learned	Recall	Precision
NT0	4.61	84.53
NT5	1.58	93.44
NT7	1.36	87.14
NT4	1.09	75.10
NT10	0.82	84.54

(b) $DA_{cluster}$, top 5 noun-matches of 135

Learned	Recall	Precision
NT0	6.93	87.09
NT4	6.48	89.91
NT8	2.62	40.48
NT11	0.86	68.60
NT10	0.58	16.95

Table 3: The top five expression classes to match N (noun) in ICE-GB, ranked by recall.

Table 2, summarizes the success of the algorithm at discovering different types of constituent. Note that these results are unlabeled, so we are examining the proportion of each type of constituent in ICE-GB that has been identified. Here, Directed Alignment exhibits the most success at identifying non-clauses, of which the primary source of success is short sentence fragments. Around 10% of noun-phrases (NP), verb-phrases (VP) and subordinate-phrases (SUBP) were recovered, this limited success reflects the nature of the constituents: all three have relatively simple constructions, whereby a single word represents the constituent. In contrast, constituent types that comprise multiple units, such as prepositional-phrases (PP), are seldom recovered.

4.3.1 Class generalization

During learning in $DA_{cluster}$, we induce generalized classes using the expression clustering algorithm. This generalization can be evaluated, comparing induced classes with those in the treebank using precision and recall. Table 2(a) shows the top five proposed classes matching the type noun (N) in ICE-GB during 500 iterations of context-sensitive Directed Alignment. There are 271 types matching noun, and as can be seen, the top five account for a very small proportion of all nouns, some 9.46% (recall).

Table 2(b) shows the same analysis for Directed Alignment with class generalization. For noun matches, we can see that there are far fewer proposed classes (135), and that those classes are much more probable, the top five accounting for 17.47%

(a) Noun Phrases (frequency=123870)

LEFT		START		END		RIGHT	
SYMB	REC	SYMB	REC	SYMB	REC	SYMB	REC
PREP	0.36	ART	0.29	N	0.53	PUNC	0.36
V	0.19	PRON	0.29	PRON	0.19	V	0.18
#STA#	0.12	N	0.2	N_2	0.11	AUX	0.13
CONJ	0.11	N_1	0.06	PUNC	0.06	CONJ	0.09
PUNC	0.09	ADJ	0.06	NUM	0.04	PREP	0.07

(b) Verb Phrases (frequency=50693)

Left		Start		End		Right	
SYMB	REC	SYMB	REC	SYMB	REC	SYMB	REC
PRON	0.32	V	0.68	V	0.98	PREP	0.20
N	0.26	AUX	0.29	PUNC	0.01	ART	0.16
PTCL	0.11	AUX_1	0.02	AUX	0.00	PRON	0.14
PUNC	0.06	V_1	0.00	V_2	0.00	ADV	0.13
CONJ	0.05	ADV	0.00	ADV	0.00	ADJ	0.09

(c) Prepositional Phrases (frequency=45777)

Left		Start		End		Right	
SYMB	REC	SYMB	REC	SYMB	REC	SYMB	REC
N	0.46	PREP	0.96	N	0.63	PUNC	0.56
V	0.23	PREP_1	0.02	N_2	0.12	CONJ	0.09
ADV	0.05	ADV	0.01	PUNC	0.08	PREP	0.09
PUNC	0.05	NUM	0.00	PRON	0.05	V	0.07
ADJ	0.04	ADV_1	0.00	NUM	0.03	AUX	0.05

Table 4: The five most frequent left/start/end/right POS contexts for NP, VP and PP constituents.

of nouns in ICE-GB. The algorithm seems to be achieving some worthwhile generalization, which is reflected in a slight increase in EVALB scores for $DA_{cluster}$. However, this increase is not a significant one, suggesting that this generalization is not sufficient to support distributional learning. We might expect this: attempting to cluster based on the low-frequency and polysemous words in expressions seems likely to produce unreliable clusters.

5 A closer look at distributional contexts

The results discussed so far seem discouraging for the approach. However, there are good reasons why these results are so poor, and why we can expect little improvement in the current formulation. We can show some of these reasons by examining actual constituent-context distributions.

Table 4 shows an analysis of the constituent types NP, VP and PP in ICE-GB, against the five most frequent POS tags[3] occurring as left-context, constituent-start, constituent-end, and right-context. We distinguish the following POS categories as being primarily functional, as they account for the majority of context-units considered by Directed Alignment: prepositions (PREP), articles (ART), auxiliaries (AUX), sentence-starts (#STA#), pronouns (PRON), conjunctions (CONJ), particles (PTCL) and punctuation (PUNC).

From Table 4, we can see that noun-phrases and verb-phrases are relatively well-suited to our approach. First, both types have strong functional left- and right-contexts: 58% of NP left-contexts and 50% of NP right-contexts are members of our functional POS; similarly, 43% of VP left-contexts and 49% of VP right-contexts are functional. This means that a probability-based model of context, such as ours, will find relatively strong support for these types. Second, both NP and VP have minimal unit types: nouns and pronouns for NP; verbs for VP. As a consequence, these types tend to carry more probability mass, since shorter sequences tend to be more frequent. We should expect our system to perform reasonably on NP and VP as a result.

In contrast, prepositional-phrases are much less amenable to distributional analysis. First, PP tend to be longer, since they contain NP, and this has obvious repercussions for alignment probabilities. More damagingly, PP contexts are dominated by open-class words - the top 74% of PP left-contexts are nouns, verbs and adverbs. Therefore, a purely probabilistic distributional approach cannot account for prepositional-phrases, since learning data is too sparse. Previous approaches have relied upon open-class generalization to reduce this problem, but these methods suffer from the same problems of data sparsity, and as such are not reliable enough to resolve the issue.

6 Attachment

We have seen that strictly probabilistic distributional analysis is not sufficient to learn constituents from raw text. If we are to improve upon this, we must find a way to identify constituents from their component parts, as well as by contextual analysis. The constituent-context distributions in Table 4 give us some clues as to where to start: both noun-phrases and prepositional-phrases show very significant constituent-starts, with articles and pronouns starting 58% of NP, and prepositions starting 94% of all PP. These functional types would be identified as contexts in Directed Alignment, but the strong relation to their containing constituents would be ig-

[3] The same trends can be shown for words, but a POS analysis is preferred for clarity and brevity.

nored.

One method for achieving such an internal relationship might be to attach contexts to the expressions with which they co-occur, and we propose using such a method here. However, this requires that we have some criterion for deciding when and how expressions should be attached to their contexts. We use a measure based on STOP arguments (Collins, 1999), which allows us to condition the decision to insert a constituent boundary on the evidence we see for doing so. For raw text, the only boundaries that are explicitly marked are at the start and end of sentences, and it is this information we use to decide when to attach contexts to expressions[4]. In other words, if a context is likely to start a sentence, we assume it is also likely to start a constituent at other positions within a sentence.

In order to calculate the likelihood of a particular context word w occurring at the start or end of a sentence, we simply use the bigram probabilities between w and the special symbols START and END, which denote the start and end of a sentence respectively. From these probabilities, we calculate Mutual Information $MI(START, w)$ and $MI(w, END)$. We prefer MI because it describes the strength of the relation between w and these special symbols without bias towards more probable words. From these MI values, we calculate a *Directional Preference* (DP) for the context word:

$$dp(w) = MI(w, END) - MI(START, w) \quad (3)$$

This yields a number representing whether w is more likely to start or end a sentence. This number will be zero if we are equally likely to see w at the start or end of a sentence, negative if w is more likely to start a sentence, and positive if w is is more likely to end a sentence.

System	UP	UR	F_1	CB
DA_{STOP}	33.6	14.1	19.8	0.42

Table 5: EVALB results after 500 iterations of Directed Alignment with STOP attachment applied to ICE-GB (DA_{STOP}).

Category	Frequency	Recall (%)
NP	117776	18.11
VP	50280	9.78
PP	42134	18.19
CL	28641	2.97
SUBP	7474	12.82
NONCL	1919	22.62

Table 6: Constituent retrieval results for DA_{STOP}, categorized by gold-type

Using DP, we can decide how to attach an expression to its contexts. For a given alignment, we consider the possibility of attaching the expression to neither context, the left-context, or the right-context, by comparing the DP for the left- and right-contexts. If the left-context shows a strong tendency to start sentences, and the right-context does not show a strong tendency to end sentences (i.e. there is an overall DP is negative), we attach the expression to its left-context; if the reverse situation is true, we attach the expression to its right context. Should the difference between these DP fall below a threshold, neither context is preferred, and the expression remains unattached.

Let us consider a specific example of attachment. The first alignment considered by the system (when applied to ICE-GB) is:

the NT1 of

Here, we need to compare the likelihood of seeing a constituent start with ‘the’ with with the likelihood of seeing a constituent end with ‘of’. Intuitively, ‘the’ occurs frequently at the start of a sentence, and never at the end. Consequently, it has a high negative DP. Meanwhile ‘of’ has a small negative DP. In combination, there is a high negative DP, so we attach the expression to the left-context, ‘the’.

6.1 Experimental Analysis

We applied Directed Alignment with attachment based on STOP arguments (DA_{STOP}) to ICE-GB as before, running for 500 iterations. These results are shown in Table 5. The results are encouraging. Unlabeled precision increased by almost 50%, from 23.6% for $DA_{cluster}$ to 33.6%. Likewise, system recall increased dramatically, from 8.1% to 14.1%, up some 75%. Crossing-brackets increased slightly, but remained relatively low at 0.42.

Table 6 shows the breakdown of EVALB scores

[4] For this method to work, we assume that our corpus is segmented into sentences. This is not the case for speech, but for learning from text it seems a reasonable assumption.

for the major non-terminal types, as before. The improvement in EVALB scores is attributable to a marked increase in success at identifying prepositional-phrases, with a lesser increase in noun-phrase identification.

6.2 Discussion

The attachment procedure described above is more successful at discovering nested constituents than distributional methods. There are good reasons why this should be the case. First, attachment compresses the corpus, removing the bias towards shorter sequences. Indeed, the algorithm seems capable of retrieving complex constituents of up to ten words in length during the first 500 iterations.

Second, the STOP-conditioning criterion, while somewhat *ad hoc* in relation to distributional methods, allows us to assess where constituent boundaries are likely to occur. As such, this can be seen as a rudimentary method for establishing argument relations, such as those observed in (Klein, 2005, Chapter 6).

Despite these improvements, the attachment process also makes some systematic mistakes. Some of these may be attributed to discrepancies between the syntactic theory used to annotate the treebank and the attachment process. For example, verbs are routinely attached to their subjects before objects, contradicting the more traditional interpretation present in treebanks. Some of the remaining mistakes can be attributed to the misalignment, due to the orthographic match problem described in Section 3.

7 Future Work

The major problem when applying distributional methods to raw text is that of orthographic matching, which causes misalignments between alternative senses of a particular word-form. To reduce this problem, context-units must be classified in some way to disambiguate these different senses. Such classification could be used as a precursor to alignment in the system we have described.

In addition, to better evaluate the quality of attachment, dependency representations and treebanks could be used, which do not have an explicit order on attachment. This would give a more accurate evaluation where subject-verb attachment is concerned.

8 Conclusions

We have presented an incremental grammar induction system that uses heuristics to improve the efficiency of distributional learning. However, in tests over a large corpus, we have shown that it is capable of learning only a small subset of constituent structure. We have analyzed actual constituent-context distributions to explain these limitations. This analysis provides the motivation for a more structured learning method, which incorporates knowledge of verifiable constituent boundaries - the starts and ends of sentences. This improved system performs significantly better, with a 75% increase in recall over distributional methods, and a significant improvement at retrieving structures that are problematic for distributional methods alone.

References

Pieter Adriaans. 1999. Learning shallow context-free languages under simple distributions. Technical Report PP-1999-13, Institute for Logic, Language, and Computation, Amsterdam.

Alexander Clark. 2001. Unsupervised induction of stochastic context free grammars with distributional clustering. In *Proceedings of the Fifth Conference on Natural Language Learning*, pages 105–112, Toulouse, France, July.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Dan Klein. 2005. *The Unsupervised Learning of Natural Language Structure*. Ph.D. thesis, Department of Computer Science, Stanford University, March.

Tony C. Smith and Ian H. Witten. 1993. Language inference from function words. Working Paper Series 1170-487X-1993/3, Department of Computer Science, University of Waikato, Hamilton, New Zealand, August.

Zach Solan, David Horn, Eytan Ruppin, and Shimon Edelman. 2003. Unsupervised efficient learning and representation of language structures. In R. Alterman and D. Kirsch, editors, *Proceedings of the 25th Conference of the Cognitive Science Society*, Hillsdale, NJ. Erlbaum.

Menno van Zaanen. 2000. Learning structure using Alignment Based Learning. In *Proceedings of the Third Annual Doctoral Research Colloquium (CLUK)*, pages 75–82.