Adapting a Lexicalized-Grammar Parser to Contrasting Domains

Laura Rimell and **Stephen Clark**
Oxford University Computing Laboratory
Wolfson Building, Parks Road
Oxford, OX1 3QD, UK
{laura.rimell,stephen.clark}@comlab.ox.ac.uk

Abstract

Most state-of-the-art wide-coverage parsers are trained on newspaper text and suffer a loss of accuracy in other domains, making parser adaptation a pressing issue. In this paper we demonstrate that a CCG parser can be adapted to two new domains, biomedical text and questions for a QA system, by using manually-annotated training data at the POS and lexical category levels only. This approach achieves parser accuracy comparable to that on newspaper data without the need for annotated parse trees in the new domain. We find that retraining at the lexical category level yields a larger performance increase for questions than for biomedical text and analyze the two datasets to investigate why different domains might behave differently for parser adaptation.

1 Introduction

Most state-of-the-art wide-coverage parsers are based on the Penn Treebank (Marcus et al., 1993), making such parsers highly tuned to newspaper text. A pressing question facing the parsing community is how to adapt these parsers to other domains, such as biomedical research papers and web pages. A related question is how to improve the performance of these parsers on constructions that are rare in the Penn Treebank, such as questions. Questions are particularly important since a question parser is a component in most Question Answering (QA) systems (Harabagiu et al., 2001).

In this paper we investigate parser adaptation in the context of lexicalized grammars, by using a parser based on Combinatory Categorial Grammar (CCG) (Steedman, 2000). A key property of CCG is that it is lexicalized, meaning that each word in a sentence is associated with an elementary syntactic structure. In the case of CCG this is a lexical category expressing subcategorization information. We exploit this property of CCG by performing manual annotation in the new domain, but only up to this level of representation, where the annotation can be carried out relatively quickly. Since CCG lexical categories are so expressive, many of the syntactic characteristics of a domain are captured at this level.

The two domains we consider are the biomedical domain and questions for a QA system. We use the term "domain" somewhat loosely here, since questions are best described as a particular set of syntactic constructions, rather than a set of documents about a particular topic. However, we consider question data to be interesting in the context of domain adaptation for the following reasons: 1) there are few examples in the Penn Treebank (PTB) and so PTB parsers typically perform poorly on them; 2) questions form a fairly homogeneous set with respect to the syntactic constructions employed, and it is an interesting question how easy it is to adapt a parser to such data; and 3) QA is becoming an important example of NLP technology, and question parsing is an important task for QA systems.

The CCG parser we use (Clark and Curran, 2007b) makes use of three levels of representation: one, a POS tag level based on the fairly coarse-grained POS tags in the Penn Treebank; two, a lexical category level based on the more fine-grained CCG lexical categories, which are assigned to words by a CCG su-

pertagger; and three, a hierarchical level consisting of CCG derivations. A key idea in this paper, following a pilot study in Clark et al. (2004), is to perform manual annotation only at the first two levels. Since the lexical category level consists of sequences of tags, rather than hierarchical derivations, the annotation can be performed relatively quickly.

For the biomedical and question domains we manually annotated approximately 1,000 and 2,000 sentences, respectively, with CCG lexical categories. We also created a gold standard set of grammatical relations (GR) in the Stanford format (de Marneffe et al., 2006), using 500 of the questions. For the biomedical domain we used the BioInfer corpus (Pyysalo et al., 2007a), an existing gold-standard GR resource also in the Stanford format. We evaluated the parser on both lexical category assignment and recovery of GRs.

The results show that the domain adaptation approach used here is successful in two very different domains, achieving parsing accuracy comparable to state-of-the-art accuracy for newspaper text. The results also show, however, that the two domains have different profiles with regard to the levels of representation used by the parser. We find that simply retraining the POS tagger used by the parser leads to a large improvement in performance for the biomedical domain, and that retraining the CCG supertagger on the annotated biomedical data improves the performance further. For the question data, retraining just the POS tagger also improves parser performance, but retraining the supertagger has a much greater effect. We perform some analysis of the two datasets in order to explain the different behaviours with regard to porting the CCG parser.

2 The CCG Parser

The CCG parser is described in detail in Clark and Curran (2007b) and so we provide only a brief description. The stages in the CCG parsing pipeline are as follows. First, a maximum entropy POS tagger assigns a single POS tag to each word in a sentence. POS tags are fairly coarse-grained grammatical labels indicating part-of-speech; the Penn Treebank set, used here, contains approximately 50 labels.

Second, a maximum entropy supertagger assigns CCG lexical categories to the words in the sentence. Lexical categories can be thought of as fine-grained POS tags expressing subcategorization information, i.e. information about the argument frame of the word. There are 425 categories in the set used by the CCG parser. Supertagging was originally developed for Lexicalized Tree Adjoining Grammar (Bangalore and Joshi, 1999), but has been particularly successful for wide-coverage CCG parsing (Clark and Curran, 2007b). Rather than assign a single category to each word, the supertagger operates as a multi-tagger, sometimes assigning more than one category if the context is not sufficiently discriminating to suggest a single tag (Curran et al., 2006). Since the taggers have linear time complexity, the first two stages can be performed extremely quickly.

Finally, the parsing stage combines the lexical categories, using a small set of combinatory rules that are part of the grammar of CCG, and builds a packed chart representation containing all the derivations which can be built from the lexical categories. The Viterbi algorithm efficiently finds the highest scoring derivation from the packed chart, using a log-linear model to score the derivations. The grammar and training data for the newspaper version of the CCG parser are obtained from CCGbank (Hockenmaier and Steedman, 2007), a CCG version of the Penn Treebank.

The aspect of the pipeline which is most relevant to this paper is the supertagging phase. Figure 1 gives an example sentence from each target domain, with the CCG lexical category assigned to each word shown below the word, and the POS tag to the right. Note that the categories contain a significant amount of grammatical information, in particular subcategorization information. The verb *acts* in the biomedical sentence, for example, looks for a prepositional phrase (PP, *as a linkage protein*) to its right and a noun phrase (NP, *Talin*) to its left, with the resulting category a declarative sentence (S[dcl]).

Bangalore and Joshi (1999) refer to supertagging as *almost parsing*, because once the correct lexical categories have been assigned, the parser is left with much less work to do. The CCG supertagger is not able to assign a single category to each word with extremely high accuracy — hence the need for it to operate as a multi-tagger — but even in multi-tagger mode it dramatically reduces the ambiguity passed through to the parser (Clark and Curran, 2007b).

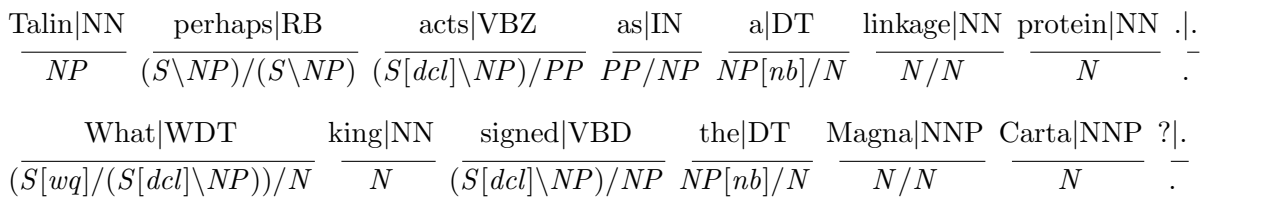

Figure 1: Example sentences with lexical category assignment.

The parser has been evaluated on DepBank (King et al., 2003), using the GR scheme of Briscoe et al. (2006), and it scores 82.4% labelled precision and 81.2% labelled recall overall (Clark and Curran, 2007a). Section 4.4 describes how the CCG dependencies can be mapped into the Stanford GR scheme (de Marneffe et al., 2006) and gives the results of evaluating the parser on biomedical and question GR resources.

The CCG parser is particularly well suited to the biomedical and question domains. First, use of CCG allows recovery of long-distance dependencies. In the sentence *What does target heart rate mean?*, the word *What* is an underlying object of the verb *mean*. The parser recovers this information despite the distance between the two words. This capability is crucial for question parsing, and also useful in the biomedical field for extraction of relationships between biological entities. Additionally, the speed of the parser (tens of sentences per second) is useful for the large volumes of biomedical data that require processing for biomedical text mining.

3 Approach

Our approach to domain adaptation is to target the coarser-grained, less syntactically complex, levels of representation used by the parser, and to train new models with manually annotated data at these levels. The motivation for this approach is twofold. First, accuracy at each stage of the pipeline depends on accuracy at the earlier stages. If the POS tagger assigns incorrect tags, it is unlikely that the supertagger will be able to recover and produce the correct lexical categories, since it relies heavily on POS tags as features. Without the correct categories, the parser in turn will be unable to find a correct parse.

In the sentence *What year did the Vietnam War end?*, the newspaper-trained POS tagger incorrectly assigns the POS tag NN (common noun) to the verb *end*, since verb-final sentences are atypical for the PTB. As a result, the supertagger is virtually certain (greater than 99% probability) that the correct CCG lexical category for *end* is N (noun). The parser then assigns *the Vietnam War end* the structure of a noun phrase, and chooses an unusual subcategorization frame for *did* in which it takes three arguments: *What*, *year*, and *the Vietnam War end*.

In the sentence *How many siblings does she have?*, on the other hand, the supertagger assigns an incorrect category to the word *How* despite it having the correct POS tag (WRB for wh-adverb). The correct category is $((S[wq]/(S[q]/NP))/N)/(NP/N)$, which takes *many* (category NP/N) and *siblings* (category N) as arguments. Instead it is tagged as $S[wq]/S[q]$, the category for a sentential adverb (i.e. the manner reading of *how*), which prevents a correct parse. Our intention was that creating new training data at the lower levels of representation would improve the accuracy of the POS tagger and supertagger in the target domains, thereby improving the accuracy of later stages in the pipeline as well.

The second motivation for our approach is to reduce annotation overhead. Full syntactic derivations are costly to produce by hand. POS tags, however, are relatively easy to annotate; even an out-of-domain tagger will provide a good starting point, and manual correction is quick, especially in a domain without much unfamiliar vocabulary. CCG lexical categories require more expertise, but our experience shows that an out-of-domain supertagger can again provide a starting point for correction, and since the annotation is flat rather than hierarchical, we hypothesize that it is not as difficult or time-consuming as annotation of full derivations.

Our adaptation approach has been partially explored in previous work which targets one or another of the different levels of representation.

Lease and Charniak (2005) obtained an improvement in the accuracy of the Charniak (2000) parser, as well as POS tagging accuracy, when applied to the biomedical domain, by training a new POS tagger model with a combination of newspaper and biomedical data. The parser improvement was due solely to the new POS tagger, without retraining the parser model. Since the Charniak parser does not use a lexicalized grammar with an intermediate level of representation, any further improvements would have to come from the parser model itself.

Clark et al. (2004) obtained an improvement in CCG supertagging accuracy for *What*-questions by training a new supertagger model with a combination of newspaper and question data annotated with CCG lexical categories. Because a question resource annotated with GRs was not available, they did not perform a parser evaluation, and the effects of the POS tagging level were not compared to the lexical category level. In this paper, we extend the pilot experiments performed by Clark et al. (2004) in four ways. First, we use a larger corpus of TREC questions covering additional question types, thus extending the experiments to the question domain more broadly, as well as to the biomedical domain. Second, we create a gold standard GR resource enabling a full parser evaluation on question data. Third, we show that the POS level is important for adaptation, reinforcing the work of Lease and Charniak (2005). A key finding of the present paper is that the combination of retraining at the POS tag and lexical category levels provides additional improvements beyond those gained by retraining at a single level. Finally, we provide analysis comparing the adaptation methodology for question and biomedical data.

Hara et al. (2007) followed a similar approach to Clark et al. (2004), using the parser of Ninomiya et al. (2006), a version of the Enju parser (Miyao and Tsujii, 2005). Enju is based on HPSG, a lexicalized grammar formalism. They obtained an improvement in parsing accuracy in the biomedical domain by training a new probabilistic model of lexical entry assignments on a combination of newspaper and biomedical data without changing the original newspaper-trained parsing model. Hara et al. (2007) did not consider the role of POS tagging. The lexical category data in Hara et al. (2007) was derived from a gold standard treebank, while the annotation of lexical categories in this paper was performed without reference to gold standard syntactic derivations.

Judge et al. (2006) produced a corpus of 4,000 questions annotated with syntactic trees, and obtained an improvement in parsing accuracy for Bikel's reimplementation of the Collins parser (Collins, 1997) by training a new parser model with a combination of newspaper and question data. Our approach differs in retraining only at the levels of representation below parse trees.

4 Experiments and Results

4.1 Resources

We have used a combination of existing resources and new, manually annotated data. The baseline POS tagger, supertagger, and parser are trained on WSJ Sections 02-21 of CCGbank. The baseline performance at each level of representation is on WSJ Section 00 of CCGbank, which contains 1913 sentences and approximately 45,000 words.

For the biomedical domain, we trained the POS tagger on gold-standard POS tags from GENIA (Kim et al., 2003), a corpus of 2,000 MEDLINE abstracts containing a total of approximately 18,500 sentences and 440,000 words. We also annotated the first 1,000 sentences of GENIA with CCG lexical categories. This set of 1,000 sentences, containing approximately 27,000 words, was used for POS tagger evaluation and for development and evaluation of a new supertagger model. For parser evaluation, we used BioInfer (Pyysalo et al., 2007a), a corpus of MEDLINE abstracts (on a different topic from those in GENIA) containing 1,100 sentences, and with syntactic dependencies encoded as grammatical relations in the Stanford GR format. We used the same evaluation set of 500 sentences as in Pyysalo et al. (2007b), and the remaining 600 for development of the mapping to Stanford format. Two parsers have already been evaluated on BioInfer, which makes it a useful resource for comparative evaluation.

For the question domain, we extended the dataset described in Clark et al. (2004). That dataset contained 1,171 questions beginning with the word *What*, from the TREC 9-12 competitions (2000-2003), manually POS tagged and annotated with

CCG lexical categories. We annotated all the additional TREC question types and improved the existing annotation, for a total of 1,828 sentences. We additionally annotated a random subset of 500 of these with GRs in the Stanford format. This subset served as our evaluation set at all levels of representation. It contains approximately 4,000 words, fewer than the other domains because of the significantly shorter sentence lengths of typical questions. The remaining 1,328 sentences were used as training data. A set of about a dozen sentences from the evaluation and training sets were used to develop the mapping to Stanford format for lexical categories not occurring in the biomedical data.

4.2 POS tagger

We began by training new models at the POS tag level of representation. All datasets use the PTB tagset. As a baseline, we used the original WSJ 02-21 model on the biomedical and question datasets. For comparison we also evaluated on Section 00 using the WSJ-trained model.

For the question data, the new POS tagger was trained on CCGbank Sections 02-21 plus ten copies of the 1,328 training sentences. The WSJ data provides additional robustness and wide grammatical coverage, and the weighting factor of ten was chosen in preliminary experiments to prevent the newspaper data from "overwhelming" the question data. For the biomedical data, the new POS tagger was trained on the full GENIA corpus, minus the first 1,000 sentences. GENIA is large enough that combination with the newspaper data was not needed.

Table 1 gives the results. For both of the new domains the performance of the WSJ model decreased compared to Section 00, but the retrained model performed at least as well as the WSJ model did on 00.[1] Improving the POS tagger performance has a positive effect on the performance of the supertagger and parser, which will be discussed in Sections 4.3-4.4.

[1]Since GENIA does not use the proper noun tag, NNP, for names of genes and other biomedical entities, all figures in this paper collapse the NNP-NN distinction where relevant for biomedical data. The question data uses NNP and the distinction is not collapsed.

	WSJ 02-21	Retrained
Sec. 00	96.7	—
Qus	92.2	97.1
Bio	93.4	98.7

Table 1: POS tagger accuracy (%) for original and retrained models.

	Orig pipeline	Retrained POS	Retrained POS and super
Sec. 00	91.5	—	—
Qus	71.6	74.0	92.1
Bio	89.0	91.2	93.0

Table 2: Supertagging accuracy (%) and the effect of retraining the POS model and the supertagger model.

4.3 Supertagger

We next trained new models at the CCG lexical category level. The training data consisted of manually annotated biomedical and question sentences; specifically, lexical categories were automatically assigned by the original parsing pipeline and then manually corrected. Whenever possible we used categories from the parser's original set of 425, although occasionally it was necessary to use a new category for a syntactic construction not occurring in CCGbank Sections 02-21. (The parser can be configured to recognize additional categories.) Question data in particular requires the use of categories that are rare or unseen in CCGbank.

For the questions, the new supertagger model, like the POS tagger, was trained on WSJ 02-21 plus ten copies of the 1,328 training sentences. For the biomedical data, a ten-fold cross-validation was performed, training each supertagger model on WSJ 02-21 plus ten copies of 90% of the 1,000 annotated sentences. Table 2 gives the supertagger accuracy with and without the retrained POS and supertagger models. The figure for the retrained biomedical supertagger is the average of the ten-fold split.

The results show an improvement in accuracy of lexical category assignment solely from retraining the POS tagger, and an additional improvement from retraining the supertagger. Supertagger accuracy for the two domains with a retrained supertagger was comparable, and in both cases was at least as high

What car company invented the Edsel?
(nsubj invented company)
(det Edsel the)
(dobj invented Edsel)
(det company What)
(nn company car)

Figure 2: Example of grammatical relations in the Stanford grammatical relation format.

as for the original pipeline on Section 00. The question data started from a much lower baseline figure, however.

4.4 Parser

We evaluated the parser on the 500 questions annotated with Stanford GRs and on the 500 evaluation sentences from the BioInfer corpus. We used the original newspaper pipeline, a pipeline with a retrained POS tagger, and a pipeline with both a retrained POS tagger and supertagger.

In order to perform these evaluations we developed a mapping from the parser's native CCG syntactic dependencies to GRs in the Stanford format. The mapping was based on the same principles as the mapping that produces GR output in the style of Briscoe et al. (2006). These principles are discussed in detail in Clark and Curran (2007a); in summary, the argument slots in the CCG dependencies are mapped to argument slots in Stanford GRs, a fairly complex, many-to-many mapping. An additional post-processing script applies some manually developed rules to bring the output closer to the Stanford format. Figure 2 gives an example of Stanford GRs, where the label of the relation is followed by two arguments, head and dependent.

Table 3 gives the results of the parser evaluation on GRs. Since the parser model was not retrained, the improvements in accuracy are due solely to the new POS and supertaggers. The results are given as an F-score over labelled GRs.[2]

The F-scores given in Table 3 are only for sentences for which a parse was found. However, there were also improvements in coverage with the retrained models. For the question data, parser coverage was 94% for the original pipeline and the pipeline with just the retrained POS tagger, and 99.6% with the retrained POS and supertaggers. For the biomedical data, coverage was 97.2% for the original pipeline, 99.0% for the pipeline with the retrained POS tagger, and 99.8% for the pipeline with the retrained POS and supertaggers.

[2]Only GRs at the lowest level of the Stanford hierarchy were considered in the evaluation; more generic relations such as *dependent* were not considered.

	Orig POS and super	New POS	New POS and super
Qus	64.4	69.4	86.6
BioInfer	76.0	80.4	81.5

Table 3: Parser F-score on grammatical relations and the effect of retraining the POS and supertagger models.

The final accuracy for both domains is in the same range as that of the original parser on newspaper data (81.8%) (Clark and Curran, 2007b), although the results are not directly comparable, since the newspaper resource uses a different GR scheme. For the BioInfer corpus, the final accuracy is also in line with results reported in the literature for other parsers (Pyysalo et al., 2007b). (No comparable GR results are available for questions.) A score in this range is thought to be near the upper bound when evaluating a CCG parser on GRs, since some loss is inherent in the mapping to GRs (Clark and Curran, 2007a).

5 Analysis

Although domain adaptation was successful for both of our target domains, the impact of the different levels of representation on parsing accuracy was not uniform. Table 3 shows that retraining the POS tagger accounted for a greater proportion of the improvement on biomedical data, while retraining the supertagger accounted for a much greater proportion on questions. In this section we discuss some of the differences between the domains which may have contributed to their behaviour in this regard, with the intention of highlighting attributes that may be relevant for domain adaptation in general.

Informally, we believe that the main difference between newspaper and biomedical text is vocabulary, and that their syntactic structures are essentially similar (with some isolated exceptions, such as more frequent use of parentheses and comma-separated

	Tag	Errors	Freq confused
Qus	WDT	129	WP
	VB	46	NN, VBP
	NNP	33	JJ, NN
	NN	32	JJ, NNP
Bio	NN	801	JJ, CD
	JJ	268	NN, VBN
	VBN	113	JJ, VBD
	FW	95	NN, IN

Table 4: Tags with the most frequent errors by the newspaper-trained POS tagger and the tags they were most frequently confused with.

lists in biomedical text). Once the POS tagger had been retrained for biomedical text, accounting for unfamiliar vocabulary, the original supertagger already performed well. The main difference between newspaper and question data, on the other hand, is syntactic. Retraining the POS tagger for questions therefore had less effect; even with the correct POS tags the supertagger was unable to assign the correct lexical categories. Since lexical categories encode syntactic information, the domain with the more divergent syntax is likely to benefit most from new training data at the lexical category level.

5.1 POS tagger

Table 1 showed that the accuracy of the newspaper-trained POS tagger was in the same range for both biomedical and question data. However, the distribution of errors was different. Table 4 shows the tags with the most frequent errors, accounting for about 75% of all POS tag errors in each domain, and the tags that they were most frequently confused with.

For the question data, the most frequent error was tagging a wh-determiner (WDT) as a wh-pronoun (WP). A determiner combines with a noun to form a noun phrase, as in the sentence *What Liverpool club spawned the Beatles?*. A pronoun, on the other hand, is a noun phrase in its own right, as in *What are the colors of the German flag?*. This tagger error arises from the fact that the word *What* occurs only once in WSJ 02-21 with a WDT tag. The second most common error was on bare verbs (VB), because the newspaper model gives a low probability of bare verbs occurring in sentence-final position, or not directly following an auxiliary.

	Unknown word rate	Unknown word-POS rate
Sec. 00	3.8	4.4
Qus	7.5	8.3
Bio	23.6	25.3

Table 5: Unknown word rate and word-POS tag pair rate (%) compared to WSJ 02-21 (by token).

For the biomedical data, the most frequent errors by far were confusions of noun (NN) and adjective (JJ). This is most likely due to the prevalence of long noun phrases in the biomedical data, such as *major histocompatibility complex class II molecules*. Although the words preceding the head noun are recognized as nominal modifiers, the classification into noun and adjective is difficult, especially when the word is previously unseen. There were also problems distinguishing verbal past participles (VBN) from adjectives (JJ) and identifying foreign words (FW), for example the phrase *in vitro*.

The fact that the newspaper-trained POS tagger performed comparably in the two target domains (Table 1) is surprising, since their lexical profiles are quite different. Lease and Charniak (2005) discussed unknown word rate as a predictor of POS tagger accuracy. However, the unknown word rate compared with WSJ 02-21 is much higher for the biomedical data than for the question data, as seen in Table 5. (The unknown word rate for the question data is still higher than that for WSJ 00, which may be due to the high proportion of proper nouns in the question data.)

Some POS tagging errors can be attributed, not to an unknown word, but to the use of a known word with an unfamiliar tag (as in the WDT example above). However, it is not the case that the question data contains many known words with unknown tags, since the rate of unknown word-tag pairs is also much higher for biomedical than for question data, as seen in the rightmost column of Table 5.

We do know that the newspaper-trained POS tagger performs better on unknown words for biomedical (84.7%) than for question data (80.4%). We hypothesize that the syntactic context of the biomedical data, being more similar to newspaper data, provides more information for the POS tagger in

	WSJ 02-21	New training sets
3-grams		
Sec. 00	0.4	—
Qus	3.6	0.7
Bio	0.7	0.5
5-grams		
Sec. 00	12.1	—
Qus	22.0	7.4
Bio	10.9	9.2

Table 6: Unknown POS n-gram rate (%) compared to WSJ 02-21, and when in-domain data is added (by token).

biomedical than in question data. Syntactic differences are discussed in the next section.

5.2 Supertagger

To quantify the syntactic distance between domains, we propose using the unknown POS n-gram rate compared to WSJ Sections 02-21. In the absence of parse trees, POS n-grams can serve as a rough proxy for the syntactic characteristics of a domain, reflecting local word order configurations. POS n-grams have been used in document modeling for text categorization (Baayen et al., 1996; Argamon-Engelson et al., 1998), but we believe our proposed use of the unknown POS n-gram rate is novel.

The leftmost column of Table 6 gives the unknown POS trigram and 5-gram rates compared to WSJ Sections 02-21. The rates for the biomedical data are quite similar to those for Section 00. The question data, however, shows higher rates of unknown POS n-grams.

For both biomedical and question data, adding in-domain data to the training set makes its syntactic profile more like that of the evaluation set. The rightmost column of Table 6 shows the unknown POS n-gram rates compared to the datasets used for training the new supertagger models, consisting of WSJ 02-21 plus annotated question or biomedical data. (For the biomedical data, the figures are averages of the same ten-fold split used for evaluation). It can be seen that adding in-domain data reduces the rate of unknown POS n-grams to about the same level observed for newspaper text.

The unknown POS n-gram rate requires POS tagged data for a new domain and thus cannot be used with unlabelled data. However, since POS tagging is relatively inexpensive, it might be possible to use this rate as one measure of syntactic distance between a training corpus and a target domain, prior to undertaking parser domain adaptation. The measure does not capture all aspects of syntactic distance, however. As pointed out by an anonymous reviewer, if the syntactic tree structures are similar across domains but lexical distributions are different – e.g. a large number of words with unfamiliar categories in the new domain – this measure will not be sensitive to the difference.

	3-grams	5-grams
Sec. 00	18	19
Qus	8	5
Bio	16	13

Table 7: Number of the 20 most frequent POS n-grams that are also in the 20 most frequent POS n-grams of WSJ Sections 02-21.

WSJ 02-21			Bio			Qus		
.	—	—	JJ	NN	NN	—	—	WP
IN	DT	NN	IN	JJ	NN	—	WP	VBZ
NN	.	—	NN	IN	JJ	—	—	WDT
DT	JJ	NN	NNS	IN	NN	WP	VBZ	DT

Table 8: Four most frequent POS trigrams for WSJ 02-21; four most frequent POS trigrams for biomedical and question data that are not in the 20 most frequent for WSJ 02-21. The dash represents the sentence boundary.

Another useful measure for comparing domain adaptation in the biomedical and question domains is frequent POS n-grams. Table 7 shows how many of the 20 most frequent POS n-grams in each dataset overlap with the 20 most frequent POS n-grams in WSJ 02-21. It can be seen that the overlap is the highest for Section 00, but much lower for the question data than for the biomedical data, again demonstrating that the question data makes frequent use of syntactic constructions which are rare in the PTB.

Table 8 shows the four most frequent POS trigrams in WSJ Sections 02-21,[3] and the four most frequent POS trigrams in the biomedical and question data that are not among the 20 most frequent

[3]Collapsing the NNP-NN distinction yields a slightly different set.

for WSJ 02-21. The frequent question trigrams include two sentence-initial question words as well as the pattern — WP VBZ, occurring in sentences beginning with e.g. *What is* or *Who is*. Though not among the top four, the pattern VB . —, representing a sentence-final bare verb, is also frequent. The most frequent biomedical POS trigrams are not dramatically different from the newspaper trigrams, but do appear to reflect the prevalence of NPs and PPs in the data.

One final measure of syntactic distance is the frequency with which CCG lexical categories that are rare or unseen in CCGbank are used in a domain. It is typical to use a few such categories, even for in-domain data, for unusual syntactic constructions, but each one is usually used only a handful of times. The question data is unique in the frequency with which previously rare or unseen categories are required. For example, the unseen category $(S[wq]/S[q])/N$, representing the word *What* in a question such as *What day did Nintendo 64 come out?* is used 11 times in the evaluation set; the rare category $(S[wq]/(S[dcl]\backslash NP))/N$, used in subject questions like *Which river runs through Dublin?*, is used 61 times; and the rare category $(S[q]/(S[pss]\backslash NP))/NP$, representing passive verbs in sentences like *What is Jane Goodall known for?*, is used 59 times.

6 Conclusion

We have targeted lower levels of representation in order to adapt a lexicalized-grammar parser to two new domains, biomedical text and questions. Although each of the lower levels has been targeted independently in previous work, this is the first study that examines both levels together to determine how they affect parsing accuracy. We achieved an accuracy on grammatical relations in the same range as that of the original parser for newspaper text, without requiring costly annotation of full parse trees.

Both biomedical and question data are domains in which there is an immediate need for accurate parsing. The question dataset is in some ways an extreme example for domain adaptation, since the sentences are syntactically uniform; on the other hand, it is of interest as a set of constructions where the parser initially performed poorly, and is a realistic parsing challenge in the context of QA systems.

Interestingly, although an increase in accuracy at each stage of the pipeline did yield an increase at the following stage, these increases were not uniform across the two domains. The new POS tagger model was responsible for most of the improvement in parsing for the biomedical domain, while the new supertagger model was necessary to see a large improvement in the question domain. We attribute this to the fact that question syntax is significantly different from newspaper syntax. We expect these considerations to apply to any lexicalized-grammar parser.

Of course, it would be useful to have a way of predicting which level of annotation would be most effective for adapting to a new domain before the annotation begins. The utility of measures such as unknown word rate (which can be performed with unlabelled data) and unknown POS n-gram rate (which can be performed with only POS tags) is not yet sufficiently clear to rely on them as predictive measures, but it seems a fruitful avenue for future work to investigate the importance of such measures for parser domain adaptation.

Acknowledgments

We would like to thank Marie-Catherine de Marneffe for advice on the use of the Stanford GR format, Sampo Pyysalo for sharing information about the BioInfer corpus, and Mark Steedman for advice on encoding question data in CCG. We would also like to thank three anonymous reviewers for their suggestions. This work was supported by EPSRC grant EP/E035698/1: Accurate and Efficient Parsing of Biomedical Text.

References

Shlomo Argamon-Engelson, Moshe Koppel, and Galit Avneri. 1998. Style-based text categorization: What newspaper am I reading? In *Proceedings of AAAI Workshop on Learning for Text Categorization*, pages 1–4.

Harald Baayen, Hans Van Halteren, and Fiona Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–131.

Srinivas Bangalore and Aravind Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the Interactive Demo Session of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL-06)*, Sydney, Australia.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Meeting of the NAACL*, pages 132–139, Seattle, WA.

Stephen Clark and James R. Curran. 2007a. Formalism-independent parser evaluation with CCG and DepBank. In *Proceedings of the 45th Meeting of the ACL*, pages 248–255, Prague, Czech Republic.

Stephen Clark and James R. Curran. 2007b. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Stephen Clark, Mark Steedman, and James R. Curran. 2004. Object-extraction and question-parsing using CCG. In *Proceedings of the EMNLP Conference*, pages 111–118, Barcelona, Spain.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Meeting of the ACL*, pages 16–23, Madrid, Spain.

James R. Curran, Stephen Clark, and David Vadas. 2006. Multi-tagging for lexicalized-grammar parsing. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL-06)*, pages 697–704, Sydney, Australia.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th LREC Conference*, pages 449–454, Genoa, Italy.

Tadayoshi Hara, Yusuke Miyao, and Jun'ichi Tsujii. 2007. Evaluating impact of re-training a lexical disambiguation model on domain adaptation of an HPSG parser. In *Proceedings of IWPT*, pages 11–22, Prague, Czech Republic.

Sanda Harabagiu, Dan Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan Bunescu, Roxana Girju, Vasile Rus, and Paul Morarescu. 2001. The role of lexico-semantic feedback in open-domain textual question-answering. In *Proceedings of the 39th Meeting of the ACL*, pages 274–281, Toulose, France.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: a corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

John Judge, Aoife Cahill, and Josef van Genabith. 2006. Questionbank: Creating a corpus of parse-annotated questions. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics*, pages 497–504, Sydney, Australia.

Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi, and Jun'ichi Tsujii. 2003. GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19:i180–i182.

Tracy H. King, Richard Crouch, Stefan Riezler, Mary Dalrymple, and Ronald M. Kaplan. 2003. The PARC 700 Dependency Bank. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora*, Budapest, Hungary.

Matthew Lease and Eugene Charniak. 2005. Parsing biomedical literature. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-05)*, Jeju Island, Korea.

Mitchell Marcus, Beatrice Santorini, and Mary Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Yusuke Miyao and Jun'ichi Tsujii. 2005. Probabilistic disambiguation models for wide-coverage HPSG parsing. In *Proceedings of the 43rd meeting of the ACL*, pages 83–90, University of Michigan, Ann Arbor.

Takashi Ninomiya, Takuya Matsuzaki, Yoshimasa Tsuruoka, Yusuke Miyao, and Jun'ichi Tsujii. 2006. Extremely lexicalized models for accurate and fast HPSG parsing. In *Proceedings of the EMNLP Conference*.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007a. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50.

Sampo Pyysalo, Filip Ginter, Veronika Laippala, Katri Haverinen, Juho Heimonen, and Tapio Salakoski. 2007b. On the unification of syntactic annotations under the stanford dependency scheme: A case study on BioInfer and GENIA. In *ACL'07 workshop on Biological, translational, and clinical language processing*, pages 25–32, Prague, Czech Republic.

Mark Steedman. 2000. *The Syntactic Process*. The MIT Press, Cambridge, MA.