Exploring Difficulties in Parsing Imperatives and Questions

Tadayoshi Hara
National Institute of Informatics, Japan
harasan@nii.ac.jp

Takuya Matsuzaki
University of Tokyo, Japan
matuzaki@is.s.u-tokyo.ac.jp

Yusuke Miyao
National Institute of Informatics, Japan
yusuke@nii.ac.jp

Jun'ichi Tsujii
Microsoft Research Asia, China
jtsujii@microsoft.com

Abstract

This paper analyzes the effect of the structural variation of sentences on parsing performance. We examine the performance of both shallow and deep parsers for two sentence constructions: imperatives and questions. We first prepare an annotated corpus for each of these sentence constructions by extracting sentences from a fiction domain that cover various types of imperatives and questions. The target parsers are then adapted to each of the obtained corpora as well as the existing query-focused corpus. Analysis of the experimental results reveals that the current mainstream parsing technologies and adaptation techniques cannot cope with different sentence constructions even with much in-domain data.

1 Introduction

Parsing is a fundamental natural language processing task and essential to various NLP applications. Recent research on parsing technologies has achieved high parsing accuracy in the same domain as the training data, but once we move to unfamiliar domains, the performance decreases to unignorable levels.

To address this problem, previous work has focused mainly on adapting lexical or syntactic preferences to the target domain, that is, on adding lexical knowledge or adjusting probabilistic models for the target domain using available resources in the target domain (see Section 2). Underlying the previous approaches, there seems to be the assumption that grammatical constructions are not largely different between domains or do not affect parsing systems, and therefore the same parsing system can be applied to a novel domain.

However, there are some cases where we cannot achieve such high parsing accuracy as parsing the Penn Treebank (PTB) merely by re-training or adaptation. For example, the parsing accuracy for the Brown Corpus is significantly lower than that for the Wall Street Journal (WSJ) portion of the Penn Treebank, even when re-training the parser with much more in-domain training data than other successful domains.

This research attempts to identify the cause of these difficulties, and focuses on two types of sentence constructions: imperatives and questions. In these constructions, words in certain syntactic positions disappear or the order of the words changes. Although some recent works have discussed the effect of these sentence constructions on parsing, they have focused mainly on more well-formed or style-restricted constructions such as QA queries, etc. This research broadens the target scope to include various types of imperatives and questions. We analyze how such sentences affect the parsing behavior and then attempt to clarify the difficulties in parsing imperatives and questions. To do so, we first prepare an annotated corpus for each of the two sentence constructions by borrowing sentences from fiction portion of the Brown Corpus.

In the experiments, parsing accuracies of two shallow dependency parsers and a deep parser are examined for imperatives and questions, as well as the accuracy of their part-of-speech (POS) tagger. Since our focus in this paper is not on the development of a new adaptation technique, a conventional supervised adaptation technique was applied to these parsers and the tagger. Our aim is rather to clarify the difficulties in parsing imperatives and questions by analyzing the remaining errors after the adaptation.

2 Related work

Since domain adaptation is an extensive research area in parsing research (Nivre et al., 2007), many ideas have been proposed, including un- or

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 749–757,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

semi-supervised approaches (Roark and Bacchiani, 2003; Blitzer et al., 2006; Steedman et al., 2003; McClosky et al., 2006; Clegg and Shepherd, 2005; McClosky et al., 2010) and supervised approaches (Titov and Henderson, 2006; Hara et al., 2007). The main focus of these works is on adapting parsing models trained with a specific genre of text (in most cases the Penn Treebank WSJ) to other genres of text, such as biomedical research papers and broadcast news. The major problem tackled in such tasks is the handling of unknown words and domain-specific manners of expression. However, parsing imperatives and questions involves a significantly different problem; even when all words in a sentence are known, the sentence has a very different structure from declarative sentences.

Compared to domain adaptation, structural types of sentences have received little attention to date. A notable exception is the work on QuestionBank (Judge et al., 2006). This work highlighted the low accuracy of state-of-the-art parsers on questions, and proposed a supervised parser adaptation by manually creating a treebank of questions.[1] The question sentences are annotated with phrase structure trees in the Penn Treebank scheme, although function tags and empty categories are omitted. QuestionBank was used for the supervised training of an LFG parser, resulting in a significant improvement in parsing accuracy. Rimell and Clark (2008) also worked on the problem of question parsing in the context of domain adaptation, and proposed a supervised method for the adaptation of the C&C parser (Clark and Curran, 2007). In this work, question sentences were collected from TREC 9-12 competitions and annotated with POS and CCG lexical categories. The authors reported a significant improvement in CCG parsing without phrase structure annotations.

Our work further extends Judge et al. (2006) and Rimell and Clark (2008), while covering a wider range of sentence constructions. Although QuestionBank and the resource of Rimell and Clark (2008) claim to be corpora of questions, they are biased because the sentences come from QA queries. For example, such queries rarely include yes/no questions or tag questions. For our study, sentences were collected from the Brown Corpus, which includes a wider range of types of questions and imperatives. In the experiments, we also used QuestionBank for comparison.

3 Target Parsers and POS tagger

We examined the performance of two dependency parsers and a deep parser on the target text sets. All parsers assumed that the input was already POS-tagged. We used the tagger in Tsuruoka et al. (2005).

3.1 MST and Malt parsers

The MST and Malt parsers are dependency parsers that produce non-projective dependency trees, using the spanning tree algorithm (McDonald et al., 2005a; McDonald et al., 2005b)[2] and transition-based algorithm (Nivre et al., 2006)[3], respectively. Although the publicly available implementation of each parser also has the option to restrict the output to a projective dependency tree, we used the non-projective versions because the dependency structures converted from the question sentences in the Brown Corpus included many non-projective dependencies. We used the `pennconverter` (Johansson and Nugues, 2007) [4] to convert a PTB-style treebank into dependency trees [5]. To evaluate the output from each of the parsers, we used the labeled attachment accuracy excluding punctuation.

3.2 HPSG parser

The Enju parser (Ninomiya et al., 2007)[6] is a deep parser based on the HPSG (Head Driven Phrase Structure Grammar) formalism. It produces an analysis of a sentence including the syntactic structure (i.e., parse tree) and the semantic structure represented as a set of predicate-argument dependencies. We used the toolkit distributed with the Enju parser to train the parser with a PTB-style treebank. The toolkit initially converts a PTB-style treebank into an HPSG treebank and then trains the parser on this. The HPSG treebank converted from the test section was used as the gold-standard in the evaluation. As evaluation metrics for the parser, we used labeled and un-

[1] QuestionBank contains a small number of imperative and declarative sentences, details of which are given in Section 4.

[2] http://sourceforge.net/projects/mstparser/

[3] http://maltparser.org/

[4] http://nlp.cs.lth.se/software/treebank_converter/

[5] We used the `-conll2007` option for the data extracted from the Brown Corpus and the `-conll2007` and `-raw` options for the QuestionBank data.

[6] http://www-tsujii.is.s.u-tokyo.ac.jp/enju

Genre	Total	Imperatives (*S-IMP*)		Questions (*SBARQ*)		Questions (*SQ*)	
		Top / embedded	Total	Top / embedded	Total	Top / embedded	Total
Popular lore	3,164	50 / 20	70 (2.21%)	40 / 11	51 (1.61%)	32 / 11	43 (1.36%)
Belles lettres	3,279	16 / 22	38 (1.16%)	51 / 9	60 (1.83%)	62 / 13	75 (2.29%)
General fiction	3,881	72 / 46	118 (3.04%)	76 / 42	118 (3.04%)	71 / 24	95 (2.45%)
Mystery / detective fiction	3,714	83 / 54	137 (3.69%)	78 / 20	98 (2.64%)	91 / 20	111 (2.99%)
Science fiction	881	17 / 18	35 (3.97%)	24 / 6	30 (3.41%)	24 / 6	30 (3.41%)
Adventure / western fiction	4,415	101 / 77	178 (1.25%)	55 / 39	94 (2.13%)	71 / 33	104 (2.36%)
Romance / love story	3,942	80 / 63	143 (3.63%)	68 / 48	116 (2.94%)	100 / 47	147 (3.73%)
Humor	967	10 / 21	31 (3.21%)	15 / 11	26 (2.69%)	25 / 18	43 (4.45%)
Total (all of the above)	24,243	429 / 321	750 (3.09%)	407 / 186	593 (2.45%)	476 / 172	648 (2.67%)

(%: ratio to all sentences in a parent genre)

Table 1: Numbers of extracted imperative and question sentences

Imperatives

- Let 's face it ! !
- Let this generation have theirs .
- Believe me .
- Make up your mind to pool your resources and get the most out of your remaining years of life .
- Believe me ! !
- Find out what you like to do most and really give it a whirl .

Questions

- Why did he want her to go to church ? ?
- Could he honestly believe it would be good for Carla to have those old prophets gripping her imagination now ? ?
- What was the matter with him that they all wearied him ? ?
- How could a man look to any one of them for an enlargement of his freedom ? ?
- Did many of Sam 's countrymen live in boxcars in the bush ? ?
- Had Sam ever lived in a boxcar ? ?

Figure 1: Example sentences extracted from Brown Corpus

labeled precision/recall/F-score of the predicate-argument dependencies produced by the parser.

4 Preparing treebanks of imperatives and questions

This section explains how we collected the treebanks of imperatives and questions used in the experiments in Section 5.

4.1 Extracting imperatives and questions from Brown Corpus

The Penn Treebank 3 contains treebanks of several genres of texts. Although the WSJ treebank has been used extensively for parsing experiments, we used the treebank of the Brown Corpus in our experiments. As the Brown Corpus portion includes texts of eight different genres of literary works (see the first column in Table 1), it is expected to contain inherently a larger number of imperatives and questions than the WSJ portion.

The Brown Corpus portion of the Penn Treebank 3 is annotated with phrase structure trees as in the Penn Treebank WSJ. Interrogative sentences are annotated with the phrase label *SBARQ* or *SQ*, where *SBARQ* denotes wh-questions, while *SQ* denotes yes/no questions. Imperative sentences are annotated with the phrase label *S-IMP*. All sentences annotated with these labels were extracted. Imperatives and questions appear not only at the top level but also as embedded clauses. We extracted such embedded imperatives and questions as well. However, if these were embedded in another imperative or question, we only extracted the outermost one. Extracted sentences were post-processed to fit the natural sentence form; that is, with first characters capitalized and question marks or periods added as appropriate.

As a result, we extracted 750 imperative sentences and 1,241 question sentences from 24,243 sentences. Examples of extracted sentences are shown in Figure 1. Table 1 gives the statistics of the extracted sentences, which show that each genre contains top-level / embedded imperative and question sentences to some extent.[7]

As described below, we also used QuestionBank in the experiments. The advantage, however, of using the Brown treebank is that it includes annotations of function tags and empty categories, and therefore, we can apply the Penn Treebank-to-HPSG conversion program of Enju (Miyao and Tsujii, 2005), which relies on function tags and empty categories. Hence, we show experimental results for Enju only with the Brown data. It should also be noted that, a constituency-to-dependency converter, `pennconverter` (Johansson and Nugues, 2007), provides a more accurate conversion when function tags and empty categories are available (see footnote 4).

4.2 Extracting questions from QuestionBank

QuestionBank consists of question sentences as well as a small number of imperative and declarative sentences. We extracted 3,859 sentences annotated with *SBARQ* or *SQ*. During the exper-

[7] Although we also applied a similar method to the WSJ portion, we only obtained 115 imperatives and 432 questions. This data was not used in the experiments.

Target	Total	Division
WSJ	43,948	39,832 (Section 02-21) for training / 1,700 (Section 22) for development test / 2,416 (Section 23) for final test
Brown overall	24,243	19,395 for training / 2,424 for development test / 2,424 for final test (randomly divided)
Brown imperatives	750	65 × 10 for ten-fold cross validation test / 100 for error analysis (chosen evenly from each genre)
Brown questions	1,240	112 × 10 for ten-fold cross validation test / 141 for error analysis (chosen evenly from each genre)
QuestionBank questions	3,859	1,000 for final test / 2,560 for training / 299 for error analysis (from the top of the corpus)

(# of sentences)

Table 2: Experimental datasets for each domain

iments, we found several annotation errors that caused fatal errors in the treebank conversion. We manually corrected the annotations of twelve sentences.[8] Examples of the annotation errors include brackets enclosing empty words and undefined or empty tags. We also found and corrected obvious inconsistencies in the corpus: character “ ’ “ replaced by “<” (737 sentences), token “?” tagged with “?” instead of “.” (2,051 sentences), and phrase labels annotated as the POS (one sentence).

5 Exploring difficulties in parsing imperatives and questions

We examined the performance of the three parsers and the POS tagger with Brown imperatives and questions, and QuestionBank questions. By observing the effect of the parser or tagger adaptation in each domain, we can identify the difficulties in parsing imperative and question sentences. We also examined the portability of sentence construction properties between two similar domains: questions in Brown and in QuestionBank.

5.1 Experimental settings

Table 2 shows the experimental datasets we created for five domains: WSJ, Brown overall, Brown imperatives, Brown questions, and QuestionBank questions. Each of the parsers and the POS tagger was adapted to each target domain as follows:

POS tagger - For Brown overall, we trained the model with the combined training data of Brown overall and WSJ. For Brown imperatives / questions and QuestionBank, we replicated the training data a certain number of times and utilized the concatenated replicas and WSJ training data for training. The number of replicas of training data was determined from among 1, 2, 4, 8, 16, 32, 64, and 128, by testing these numbers on the development test sets in three of the ten datasets for cross validation.

MST and Malt parser - For Brown overall and QuestionBank questions, we trained the model on combined data for the target domain and the original model. For Brown imperatives and questions, we replicated the training data ten times and utilized the concatenated replicas and WSJ training data for training.

Target	WSJ tagger	Adapted tagger
WSJ	97.53%	-
Brown overall	96.15%	96.68%
Brown imperatives	92.36%	93.96%
Brown questions	94.69%	95.80%
QuestionBank questions	93.14%	95.69%

Table 3: Accuracy of each POS tagging system for imperatives and questions

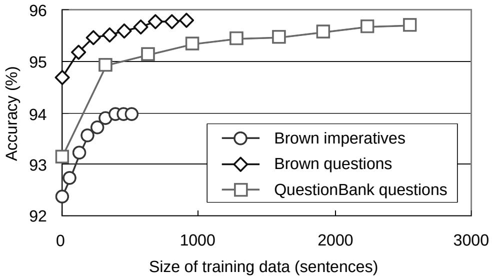

Figure 2: POS tagging accuracy vs. corpus size

Enju parser - We used the adaptation toolkit in the Enju parser (Hara et al., 2007), which is based on the idea of reference distribution (Jelinek, 1998). The parser was trained on the same training data set as the MST and Malt parser for each of the target domains .

5.2 Overview of POS tagging accuracy

Table 3 gives the POS tagging accuracy for the target domains. When we applied the WSJ tagger to other domains, the tagging accuracy basically decreased. For Brown overall, compared with the WSJ, the accuracy did not decrease much. However, for imperatives and questions, the POS tagger accuracy decreased significantly. The table shows that the adaptation improved the tagging accuracy to some extent, but that the improved accuracy for imperatives and questions was still below that of the adapted tagger for Brown overall.

Figure 2 shows the POS tagging accuracy for

[8] We intend making these corrections publicly available.

Target	Correct → Error	WSJ tagger	Adapted tagger
Brown imperatives	VB → NN / NNP	13	9
	VB → VBP	8	2
	RB → RP	4	2
	UH → RB	3	1
	RB → IN	3	1
	IN → RP	3	3
	NN → NNP	2	3
	RB → DT	2	3
Brown questions	VB → VBP	16	2
	NN → JJ	3	3
	JJ → VBN	3	3
	IN → RB	2	3
QuestionBank questions	WDT → WP	28	0
	NN → NNP	17	12
	JJ → NNP	9	7
	VB → VBP	8	1
	VB → NN / NNP	7	6
	NN → JJ	6	4

(# of errors)

Table 4: Main tagging errors for each construction

the target domains for varying sizes of the target training data. This graph shows that for both types of sentences, the first 300 training sentences greatly improved the accuracy, but thereafter, the effect of adding training data declined. It indicates the inherent difficultly in parsing imperatives and questions; to match the tagging accuracy of the WSJ tagger for the WSJ (97.53% in Table 3), just using much more training data does not appear to be enough. In particular, the problem is more serious for imperatives.

5.3 Error analysis in POS tagging

Next, we explored the tagging errors in each domain to observe the types of errors from the WSJ tagger and which of these were either solved by the adapted taggers or remain unsolved.

Table 4 shows the most frequent tagging errors given by the WSJ tagger / adapted tagger for Brown questions, Brown imperatives, and QuestionBank questions, respectively. From the results, we found that the main errors of the WSJ tagger for the Brown domains were mistagging of verbs, that is, "VB → ***". We then analyzed why each of these errors had occurred.

For Brown imperatives, the WSJ tagger gave two main tagging errors: "VB → NN(P)" and "VB → VBP". These two types of errors arise from the differences in sentence constructions between Brown imperatives and WSJ. First, the WSJ tagger, trained mainly on declarative sentences, prefers to give noun phrase-derived tags at the beginning of a sentence, whereas an imperative sentence normally begins with a verb phrase. Second, the main verb in an imperative sentence takes a base form, whereas the WSJ tagger trained mainly on tensed sentences prefers to take the verb as a present tense verb.

After adapting the tagger to Brown imperatives, the tagger would have learned that the first word in a sentence tends to be a verb, and that the main verb tends to take the base form. Table 4 shows that the above two types of errors decreased to some extent as expected, although a few mistags of verbs still remained.

By investigating the remaining errors associated with VB, we found that several errors still occurred even in simple imperative sentences such as "VB → NN" for "Charge" in "Charge something for it", and that some errors tended to occur after a to-infinitive phrase or conjunction, such as "VB → NN" for "subtract" in "To find the estimated net farm income, subtract ...". The former type could be solved by increasing the training data, whereas the latter error type cannot easily be solved with a model based on a word N-gram since it cannot recognize the existence of a long advervial phrase, etc. preceding the main verb.

We also analyzed the errors in Brown questions and QuestionBank questions, and again found that many errors were due to the fact that the WSJ tagger was trained on a corpus consisting mainly of declarative sentences. After the adaptation, although some of the errors, such as the special use of wh-words, i.e., "WDT → WP", were corrected, other kinds of errors related to the global change in sentence structure still remained.

To tag words correctly both in imperatives and questions, we may have to consider richer information than only N-gram based features, such as dependency or phrasal structures. Context information may also help; if the tagger knows that a sentence is uttered in a sequence of conversation, the tagger can consider the higher possibility of the sentence being an imperative or question.

5.4 Overview of parsing accuracy

Table 5 gives the parsing accuracy of MST (first order), MST (second order), Malt, and the Enju parser for WSJ, Brown overall, Brown imperatives, Brown questions, and QuestionBank questions. Figure 3 plots the parsing accuracy against the training data size of the four parsers for Brown imperatives, Brown questions, and QuestionBank questions. The bracketed numbers give the accuracy improvements from "WSJ parser + WSJ tagger". Note that, since the training of the MST

Parser	Target	WSJ parser + WSJ tagger	WSJ parser + Adapted tagger	WSJ parser + Gold POS	Adapted parser + WSJ tagger	Adapted parser + Adapted tagger	Adapted parser + Gold POS
MST parser (1st order)	WSJ	87.08	-	88.54 (+1.46)	-	-	-
	Brown overall	80.83	81.14 (+0.31)	82.20 (+1.37)	82.49 (+1.66)	83.00 (+2.17)	84.30 (+3.47)
	Brown imperatives	76.60	78.34 (+1.74)	81.16 (+4.56)	78.62 (+2.02)	80.86 (+4.26)	83.40 (+6.80)
	Brown questions	75.91	77.57 (+1.66)	79.83 (+3.92)	78.67 (+2.76)	80.28 (+4.37)	82.75 (+6.84)
	QuestionBank questions	59.58	60.67 (+1.09)	61.54 (+1.96)	83.25 (+23.67)	85.41 (+25.83)	86.75 (+27.17)
MST parser (2nd order)	WSJ	88.22	-	89.74 (+1.52)	-	-	-
	Brown overall	81.60	81.83 (+0.23)	83.14 (+1.54)	(-)	(-)	(-)
	Brown imperatives	76.64	78.35 (+1.71)	81.17 (+4.53)	79.44 (+2.80)	81.39 (+4.75)	84.04 (+7.40)
	Brown questions	75.92	77.65 (+1.73)	79.86 (+3.94)	(-)	(-)	(-)
	QuestionBank questions	59.63	60.60 (+0.97)	61.64 (+2.01)	(-)	(-)	(-)
Malt parser	WSJ	87.46	-	88.99 (+1.53)	-	-	-
	Brown overall	79.50	79.76 (+0.26)	80.95 (+1.45)	82.28 (+2.78)	82.59 (+3.09)	83.84 (+4.34)
	Brown imperatives	73.37	74.62 (+1.25)	77.57 (+4.20)	77.91 (+4.54)	79.92 (+6.55)	83.18 (+9.81)
	Brown questions	71.12	72.41 (+1.29)	75.25 (+4.13)	78.73 (+7.61)	80.03 (+8.91)	82.72 (+11.60)
	QuestionBank questions	58.75	59.82 (+1.07)	60.42 (+1.67)	89.28 (+30.53)	92.55 (+33.80)	93.87 (+35.12)
Enju parser	WSJ	89.56	-	90.52 (+0.96)	-	-	-
	Brown overall	81.19	81.61 (+0.42)	82.63 (+1.44)	83.70 (+2.51)	84.29 (+3.10)	85.37 (+4.18)
	Brown imperatives	74.82	76.68 (+1.86)	80.52 (+5.70)	79.91 (+5.09)	81.53 (+6.71)	84.29 (+9.47)
	Brown questions	76.88	79.45 (+2.57)	80.75 (+3.87)	80.10 (+3.22)	82.24 (+5.36)	83.55 (+6.67)

(Enju: F-score of predicate-argument relations / MST, Malt: Accuracy of labeled attachments (%))

Table 5: Accuracy of each parsing system for the Brown Corpus and QuestionBank

parser (second order) on Brown overall, Brown questions, and QuestionBank could not be completed in our experimental environment[9], the corresponding parsing accuracies denoted by bracketed hyphens in Table 5 could not be measured, and consequently, we could not plot complete graphs of the second order MST for Brown questions and QuestionBank questions in Figure 3.

After adaptation (see "Adapted parser" columns in Table 5), the parser achieved two to eight percent higher accuracy for each of the Brown domains compared to the WSJ parser. For QuestionBank, 25 to 35 percent improvement in accuracy was observed. Figure 3 shows that the improvement is generally proportional to the size of the training data and that this tendency does not seem to converge, except for the Malt parser for QuestionBank. This would suggest that lower accuracy than that of the WSJ parser for the WSJ could still be as a result of a lack of training data. In Figure 3, the parser accuracy for QuestionBank, for which we could use much more training data than for Brown questions, approaches or even exceeds that of the WSJ parser for WSJ. However, as there is no more training data for Brown imperatives and questions, we need to either prepare more training data or explore approaches that enable the parsers to be adapted with small amounts of training data.

5.5 Error analysis on parsing

To capture an overview of the adaptation effects, we observed the error reduction in the Malt parser.

Target	Dependency	WSJ parser	Adapted parser
Brown imperatives	ADV	39	32
	ROOT	33	9
	COORD	26	22
	NMOD	25	26
	OBJ	22	19
Brown questions	ADV	43	33
	NMOD	37	34
	SBJ	32	24
	ROOT	24	9
	COORD	21	12

(# of recall errors)

Table 6: Main parsing errors of Malt parser for Brown imperatives and questions

Table 6 gives the recall errors on labeled dependencies, which were observed more than ten times for 100 analysis sentences in each domain. For each dependency shown in the second column, the third and fourth columns show the number of parsing errors by the WSJ parser with gold tags and the adapted parser with gold tags, respectively. Since ROOT dependencies, that is, heads of sentences, are critical to the construction of sentences, we focus mainly on this type of error.

For Brown imperatives and questions, the reduction in ROOT dependency accuracy was prominent. On investigation, we found that the WSJ parser often made mistakes in parsing sentences which began or ended with the name of the person being addressed. For example, in Brown imperatives, for the sentence "See for yourself, Miss Zion.", the WSJ parser mistook the name "Zion" to be ROOT, and the main verb "See" to be a modifier of the name. The adapted parser correctly assigned ROOT to the main verb.

We also found that the WSJ parser often made mistakes in parsing sentences containing quota-

[9] The reason is a crash of the learning program, presumably caused by huge memory consumption and/or huge intermediate files.

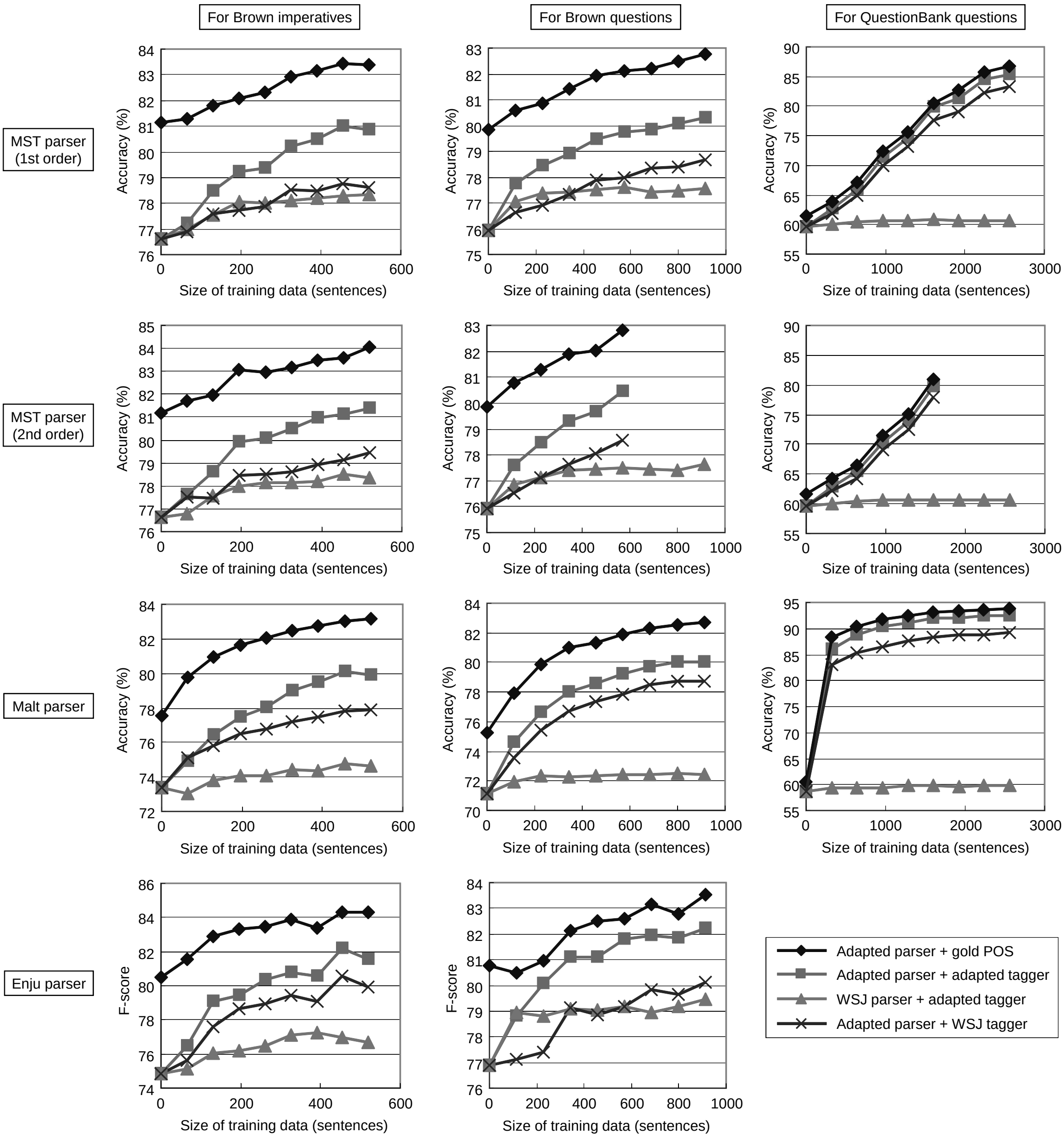

Figure 3: Learning curve of various parsers for Brown Corpus and Question Bank

tion, exclamation, or question marks, such as " "Hang on" !!" " or " "Why did you kill it" ??". For such sentences, the WSJ parser regarded the first "!" or "?" as ROOT, and "Hang" or "did" as the modifier of the punctuation. A possible reason for this type of error could be that the Brown Corpus places exclamation or question marks outside, instead of inside the quotation. The adapted parser could handle this dubious construction and assigned ROOT to the main verbs as the corpus required. [10]

On the other hand, we also observed some unsolved errors, of which we discuss two. First, Brown imperatives and questions, include many colloquial sentences, which have rather flexible constructions, especially imperatives, such as "Lift, don't shove lift!", "Come out, come out in the meadow!", etc. The parsing models based on the plausibility of constructions were not able to capture such sentences.

Second, having different sentence constructions within a single sentence, such as, where a to-infinitive phrase or subordinate clause precedes an imperative or question, often confused the parser. For example, for the imperative sentence, "To find the estimated net farm income, subtract

[10] We may have to correct the corpus.

Parser	WSJ parser + Gold POS	WSJ parser + WSJ tagger	WSJ parser + Adapted tagger	Adapted parser + Gold POS	Adapted parser + WSJ tagger	Adapted parser + Adapted tagger
Adapted to Brown questions → tested on QuestionBank questions						
MST parser (1st order)	61.54	59.58	59.65	63.06	61.07	61.07
MST parser (2nd order)	61.64	59.63	59.58	(-)	(-)	(-)
Malt parser	60.42	58.75	58.64	62.12	60.54	60.48
(POS tagger)	100	93.14	92.97	100	93.14	92.97
Adapted to QuestionBank questions → tested on Brown questions						
MST parser (1st order)	79.83	75.91	76.30	72.26	69.02	69.07
MST parser (2nd order)	79.86	75.92	76.11	(-)	(-)	(-)
Malt parser	75.25	71.12	71.15	67.70	63.98	63.43
(POS tagger)	100	94.69	94.69	100	94.69	94.69

(Parser: Accuracy of labeled attachments (%) / POS: Accuracy of tagged labels (%))

Table 7: Accuracy of each parsing system adapted to one question domain and tested on another question domain

Target	yes-no questions	wh-questions		
		WP	WDT	WRB
Brown questions	59	18	2	22
QuestionBank questions	0	48	31	21

(# of sentences in the analyzed data)

Table 8: Distribution of question types

the estimated annual farming expenditure...", both the WSJ and adapted parsers regarded "find" as ROOT, because the parsers regarded the words following "find" as a that-clause complementing "find", as in "To find [(that) the estimated net farm income, subtract the estimated annual farming ...]". It would be difficult for the parsers to know which is the main clause in such complex sentences. This type of error cannot be solved merely by increasing the training data.

Imperative or question sentences typically consist not only of a pure imperative or question clause, but also of other constructions of phrases or clauses. These complex sentences were parsed without being partitioned into separate constructions, and as a result the parser sometimes became confused.

5.6 QuestionBank vs. Brown questions

Both the Brown questions and QuestionBank are in the question domain. In this section, we examine whether a parser adapted to one domain could be ported to another domain.

QuestionBank does not provide function tags, and therefore in training and evaluation of the parsers, abstracted dependencies were extracted from the corpus. As a result, a parser adapted to one domain could not provide correct dependency labels on functions for the other domain. However, we would expect that sentence constructions are basically common and portable between two domains, which would provide a correct boundary for phrases and therefore, the correct dependencies in phrases would be introduced by the adaptation.

Table 7 gives the parsing or tagging accuracy of each parser and the POS tagger for Brown questions and QuestionBank. These results differ from those in Table 5 in that the parsers and the tagger have been adapted to another question domain. The table shows that the parsers adapted to the Brown questions improved their parsing accuracy with QuestionBank, whereas the parsers adapted to QuestionBank decreased in accuracy. Table 8 could explain this result. Using Brown questions, many wh-questions were learnt, which is what QuestionBank mainly contains. On the other hand, despite yes-no questions constituting more than half the Brown Corpus, these were not learnt using QuestionBank for training.

A question domain contains various types of questions with various sentence constructions. In order to parse questions correctly, we need to capture each of these correctly. This type of problem was not so obvious when we were working mainly with declarative sentences.

6 Conclusion

Through experiments with various parsers we observed that simple supervised adaptation methods are insufficient to achieve parsing accuracy comparable with that of declarative sentences. This observation holds both for POS tagging and parsing, and indicates that the parsers need to be fundamentally improved, such as re-constructing feature designs or changing parsing models.

Following on from this study, future work includes investigating parsing frameworks that are robust for sentences with different constructions, and/or methods that can effectively adapt a parser to different sentence constructions including imperatives and questions, among others.

References

John Blitzer, Ryan Mcdonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia.

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

A. B. Clegg and A. Shepherd. 2005. Evaluating and integrating treebank parsers on a biomedical corpus. In *Proceedings of the ACL 2005 Workshop on Software*, Ann Arbor, Michigan.

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 10th International Conference on Parsing Technologies (IWPT 2007)*, pages 11–22.

Frederick Jelinek. 1998. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, MA, USA.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for english. In *Proceedings of NODALIDA 2007*.

John Judge, Aoife Cahill, and Josef van Genabith. 2006. Questionbank: Creating a Corpus of Parsing-Annotated Questions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 497–504.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 337–344, Sydney, Australia.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic Domain Adaptation for Parsing. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL*, pages 28–36, Los Angeles, California.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005a. Online large-margin training of dependency parsers. In *Proceedings of ACL-2005*.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005b. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of HLT/EMNLP-2005*.

Yusuke Miyao and Jun'ichi Tsujii. 2005. Probabilistic disambiguation models for wide-coverage HPSG parsing. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 83–90.

Takashi Ninomiya, Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2007. A log-linear model with an n-gram reference distribution for accurate hpsg parsing. In *Proceedings of IWPT 2007*, June. Prague, Czech Republic.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, pages 2216–2219.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic, June. Association for Computational Linguistics.

Laura Rimell and Stephen Clark. 2008. Adapting a Lexicalized-Grammar Parser to Contrasting Domains. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 475–584.

Brian Roark and Michiel Bacchiani. 2003. Supervised and unsupervised PCFG adaptation to novel domains. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 126–133, Edmonton, Canada.

Mark Steedman, Miles Osborne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Bootstrapping statistical parsers from small datasets. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 331–338, Budapest, Hungary.

Ivan Titov and James Henderson. 2006. Porting statistical parsers with data-defined kernels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 6–13, New York City.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, volume LNCS 3746, pages 382–392, Volos, Greece, November. ISSN 0302-9743.