Towards Semi-Automated Annotation for Prepositional Phrase Attachment

Sara Rosenthal, William J. Lipovsky, Kathleen McKeown, Kapil Thadani, Jacob Andreas

Columbia University
New York, NY
{sara,kathy,kapil}@cs.columbia.edu, {wjl2107,jda2129}@columbia.edu

Abstract

This paper investigates whether high-quality annotations for tasks involving semantic disambiguation can be obtained without a major investment in time or expense. We examine the use of untrained human volunteers from Amazon's Mechanical Turk in disambiguating prepositional phrase (PP) attachment over sentences drawn from the Wall Street Journal corpus. Our goal is to compare the performance of these crowdsourced judgments to the annotations supplied by trained linguists for the Penn Treebank project in order to indicate the viability of this approach for annotation projects that involve contextual disambiguation. The results of our experiments show that invoking majority agreement between multiple human workers can yield PP attachments with fairly high precision, confirming that this crowdsourcing approach to syntactic annotation holds promise for the generation of training corpora in new domains and genres.

1. Introduction

The availability of training data is generally the biggest bottleneck to the performance of automated systems applied to natural language processing problems. Most standard techniques for analyzing lexical, syntactic or semantic aspects of text rely on existing annotated resources for development. Under the standard paradigm of annotation projects, the construction of new annotated corpora is expensive and requires time for training annotators to perform the task. However, Snow et al. (2008) show that using a large number of untrained annotators can yield annotations of quality comparable to those produced by a smaller number of trained annotators on multiple-choice labeling tasks. This leads us to consider whether such a crowdsourcing approach can be applied towards the collection of corpora for tasks that require semantic disambiguation.

Although the LDC[1] provides a spectrum of corpora in many languages annotated for specific tasks and representations, providing complete coverage for the vast array of domains and genres that require language processing tools is an immense challenge. However, it has been widely observed that statistical systems perform poorly when applied to text from a different domain or genre than that of their training corpora. For example, parsers trained on newswire text exhibit a clear drop in accuracy when run on weblog text and automated speech recognition systems trained on broadcast news do not perform as well on telephone conversations. This leads us to question whether various annotation paradigms can be extended to new domains without a large overhead in cost and time, and whether complex structured tasks can be achieved without trained experts.

In this work, we present an experiment on prepositional phrase (PP) attachment in order to determine whether annotators without formal linguistic training are capable of producing high-quality annotations involving semantic disambiguation. Our experiment tests whether human volunteers on Amazon's Mechanical Turk[2], an online task marketplace, can identify the correct attachment for PPs without much error when compared to the gold-standard Penn Treebank annotations provided by trained linguists. We hope to observe that errors in judgment by a single untrained human can be mitigated by the collective judgments of multiple volunteers, which can be collected quickly and at little expense.

To this end, we have developed an automated system to pose PP-attachment disambiguation tasks as multiple choice questions, a format that can be easily understood by humans unfamiliar with language processing. Our system extracts PPs and the noun or verb phrases that they attach to from Penn Treebank parse structure, along with syntactically-plausible alternatives for attachment. These multiple-choice problems were presented to workers on Amazon's Mechanical Turk and the judgments were aggregated and analyzed.

An analysis of our question-construction system shows that it yields few errors in its output. Furthermore, our evaluation of worker performance shows that using just three Mechanical Turk workers per question is sufficient for high-accuracy identification of PP-attachment, a result that confirms the viability of this semi-automated approach for the efficient annotation of corpora for similar contextual disambiguation tasks.

2. Related work

The Penn Treebank corpus (Marcus et al., 1993) has frequently been used as a source of data for a wide range of projects that require training data. It has been used for training and development in areas such as chunking (Tjong Kim Sang and Buchholz, 2000), POS tagging (Brill, 1995), and syntactic parsing (Charniak, 2000; Collins, 2003). In addition, the Wall Street Journal section has also been used as an additional annotation resource in semantic role labeling annotation for verbs in Propbank (Palmer et al., 2005), nouns in Nombank (Meyers et al., 2004), and multiple semantic annotations including named-entity recognition and word senses in Ontonotes (Pradhan and Xue, 2009). The quantity of annotated data has been tremendously useful for pushing the field forward, allowing new machine learning approaches as well as quantitative evaluations through

[1] http://www.ldc.upenn.edu
[2] https://www.mturk.com

comparison with a gold standard. Reliance on the Penn Treebank, however, means that the field is armed with tools which work well when applied to well-formed text with vocabulary similar to that found in the Wall Street Journal. These tools degrade, sometimes dramatically, when applied to data from other sources such as blogs, email, speech, or medical texts.

The performance of automated PP-attachment disambiguation systems has traditionally been evaluated on the RRR dataset (Ratnaparkhi et al., 1994), which also uses the Wall Street Journal as a data source. The dataset contains quadruples of the form {V, N1, P, N2}, where the prepositional phrase {P, N2} is attached to either V or N1. The best results using RRR achieved 81.8% accuracy (Stetina and Nagao, 1997). However, this corpus has recently come under criticism for its unrealistic simplification of the PP-attachment task that presumes the presence of an oracle to extract the two hypothesized structures for attachment (Atterer and Schütze, 2007). For these reasons, our system for finding PPs and potential attachments uses full sentences and generates more than two attachment points for each PP.

The automatic resolution of PP-attachment ambiguity is an important task which has been tackled extensively in the past (Ratnaparkhi et al., 1994; Yeh and Vilain, 1998; Stetina and Nagao, 1997; Zavrel et al., 1997). Recent work in automatic PP-attachment achieved 83% accuracy using word sense disambiguation to improve results in a parsing context (Agirre et al., 2008). While progress in this task has been steady, it is unclear whether these results would carry over to new domains and genres of text without additional training data. In this article, we explore whether crowdsourced judgments can be used to build these types of training corpora.

Amazon's Mechanical Turk has recently become a popular tool for the collection of annotated data from volunteer workers. The quality of aggregate crowdsourced judgments has been evaluated over a wide range of labeling tasks such as affect recognition, word similarity, recognizing textual entailment, event temporal ordering and word sense disambiguation; in all cases, the results were found to be in high agreement with the annotations of expert annotators (Snow et al., 2008). Our goal in this paper also involves the evaluation of worker as annotators; however, while Snow et al. (2008) manually select questions designed to evaluate worker performance, we explore a technique of automated question formulation that is targeted to the larger-scale task of corpus construction.

The ability to quickly collect syntactic and semantic annotations has significant implications for extending existing natural language processing tools to new areas. The inclusion of syntactic structure has had a major impact on tasks which involve the interpretation of text, including summarization and question answering. This leads us to hypothesize that textual analysis in noisier genres (such as emails, blogs and other webtext) could similarly be improved by annotating corpora in that genre with syntactic and semantic information like PP attachment. For example, research in areas such as information extraction (Hong and Davison, 2009), social networking (Gruhl et al., 2004), and sentiment analysis (Leshed and Kaye, 2006) could harness new datasets built for specific domains and genres using the ideas discussed in this article.

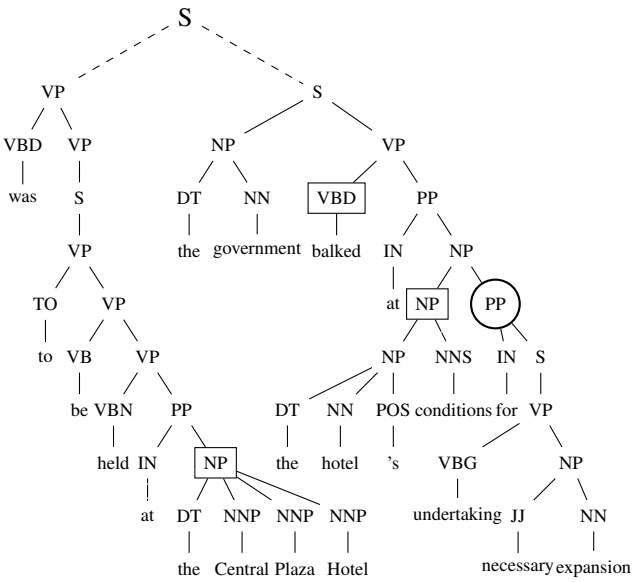

Figure 1: A partial parse of the sentence "The meeting, which is expected to draw 20,000 to Bangkok, was going to be held at the Central Plaza Hotel, but the government balked at the hotel's conditions *for undertaking necessary expansion*." from WSJ0037. The extracted PP is marked with a circle and potential attachment points are indicated with rectangles.

3. Finding PPs and their attachments

To enable workers to annotate PP attachment, we developed a system to create questions using sentences containing PPs from the Penn Treebank. These questions were multiple-choice and asked workers to choose which phrase from a set of choices was modified by the PP in question. Our system extracted sentences containing PPs from the Penn Treebank and traversed their parse trees to find each PP as well as the phrase it modifies. In general, the first sibling to the left (or first phrase to the left at this level in the tree) of the PP was considered to be the constituent the PP was modifying, accounting for punctuation. If the system found two PPs connected by a conjunction (PP CC PP), the shorter PPs were attached to the correct NP/VP. For example, in the sentence "The dog *with the spots* and *with the long tail*", the PPs "*with the spots*" and "*with the long tail*" would be attached to "The dog".

After successfully finding the correct answer, the sentence was re-examined to find additional phrases to use as plausible incorrect answers for the multiple-choice questions by looking at noun phrases and verbs that occurred in the parse prior to the PP. Three incorrect answers were created when possible; in some cases, fewer alternatives were available. Figure 1 illustrates the alternative attachment points produced using these rules for an example sentence; the question generated for this example is shown in Table 1.

The meeting, which is expected to draw 20,000 to Bangkok, was going to be held at the Central Plaza Hotel, but the government balked at the hotel's conditions for undertaking necessary expansion.
Consider the sentence above. Which of the following is the correct pairing between the prepositional phrase **for undertaking necessary expansion**, and the phrase it is meant to modify? ○ balked ○ the Central Plaza Hotel ○ the hotel 's conditions

Table 1: Multiple-choice question extracted from the parse from Figure 1

3.1. System Limitations

For this experiment, we only tested cases involving backward attachment and avoided examples such as "*In the morning*, I began to work on the paper." where the PP occurs first and modifies the verb phrase which follows. In order to maintain consistency, phrases past the PP were also not examined for the purpose of constructing incorrect answers for cases of backward attachment.

In addition, the approach described above had difficulty identifying the correct attachment when an adverb appeared in the first parse at that level of the tree; the correct attachment in these cases was the entire verb phrase which included the adverb and the verb or adjective immediately following it. We also encountered some sentences containing two identical phrases which caused duplicate correct options. In the evaluation described in section 5, these shortcomings resulted in 18 responses to questions that were mistakenly labeled as incorrect responses by the system.

4. User Studies

4.1. Pilot Study

A pilot study was carried out using Mechanical Turk prior to automatically extracting PP attachments in order to determine how to ask the questions to enable the average person to provide a good answer. 20 PP-attachment tasks were manually extracted from a single Wall Street Journal article and posed as questions in three different ways. We considered variations to examine whether the PP should be mentioned in the options for attachment as well as whether the term "prepositional phrase" could be used in the question as opposed to framing the task in more general terms. These questions were then each posed to five workers on Mechanical Turk.

From the study, we concluded that it was best to show the sentence at the top of the screen, followed by the task description, an example, and then the multiple answer choices. The final study listed attachment options without the PP and used the task description wording that yielded the most accurate attachments (16/20); this is shown in Table 1.

Attachment	Example
Adjective	She arrived **full** *of the energy and ambitions reformers wanted to reward*
Preposition	The decline was even steeper **than** *in September*
Implicit	He has as strong a presence in front of the camera as he does *behind it*
Forward	High test scores, *on the other hand*, **bring** recognition and extra money

Table 2: Examples of attachment cases that were excluded from this study; correct attachments for the italicized PPs are shown in boldface

4.2. Full Study

The question-construction system described in Section 3 was initially run on 3000 sentences with PPs from the Penn Treebank. From the resulting questions, we manually selected the first 1000 questions such that no sentence was too similar to a sentence selected earlier, the PP was not itself part of one of the answers (an error), or the correct attachment was not misidentified due to a complex sentence which produced an atypical parse tree in the Penn Treebank. Complex sentences include cases shown in Table 2, where attachment was applied to an adjective, another preposition, an intervening particle, an omitted phrase, or cases in which there was no clear backward attachment in the sentence. For future experiments, we plan on augmenting our method so that it handle cases of forward or sentence level attachments.

The full study was limited to Mechanical Turk workers who self-identified as United States residents in order to restrict the worker pool to native English speakers. Each of the 1000 questions was posed to three workers and a worker could not answer the same question twice.

Workers were paid four cents per question answered and given a maximum of five minutes to supply an answer to a question. Average completion time per task was 49 seconds and the effective hourly rate was $2.94. The entire task took 5 hours and 25 minutes to complete and the total expense was $135.00; $120 was spent on users and $15 was spent on Mechanical Turk fees. We discuss the quality of worker responses in the following section.

5. Evaluation & Analysis

Of the 3000 individual responses received by Mechanical Turk workers (three for each of the 1000 questions), 86.7% were answered correctly. On a per-question basis, the correct attachment was chosen by a majority of workers in 90.4% of the 1000 cases and unanimously in 70.8% of the cases. We manually examined all cases in which a majority of responses agreed on the wrong answer and found that in 18/96 cases, the responses were mislabeled due to the shortcomings of our system described in section 3.1. Accounting for these errors results in an improvement in worker accuracy to 92.2% for majority agreement and 71.8% for unanimous agreement.

Sentence	Prepositional phrase	Attachment options			
		Option 1	Option 2	Option 3	Option 4
'The morbidity rate is a striking finding among those of us who study asbestos-related diseases,' said Dr. Talcott	among those of us who study asbestos-related diseases	is	morbidity rate	striking finding	
The thrift holding company said it expects to obtain regulatory approval and complete the transaction by year-end.	by year end	obtain regulatory approval	complete the transaction	complete	expects

Table 3: Examples of problematic PP attachment cases

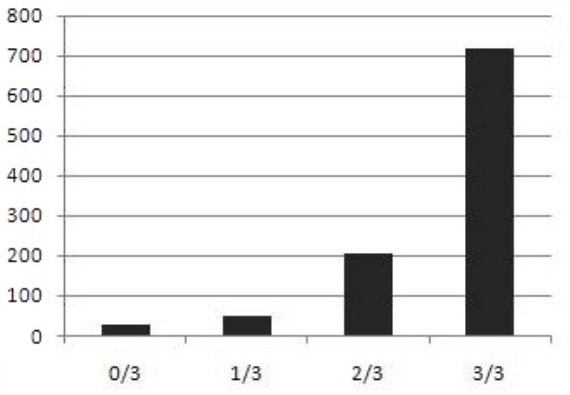

Figure 2: Number of questions answered correctly by x workers

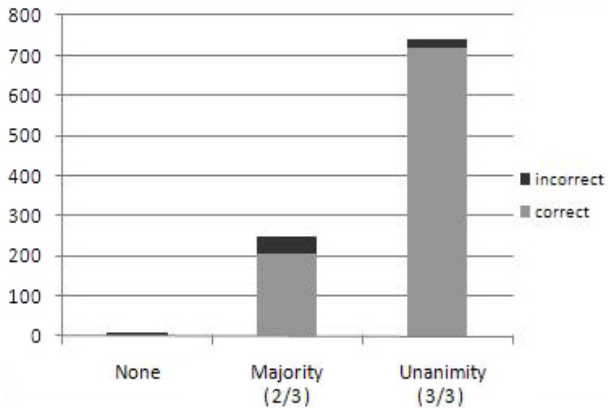

Figure 3: Number of questions in which x workers agreed upon a response

We observe that using majority opinion across just three workers on Mechanical Turk yields an accuracy that is high enough to be used reliably for corpus creation (Pradhan and Xue, 2009). These results seem to indicate that non-experts are fairly capable at disambiguating attachment for sentences in the domain of Wall Street Journal text.

5.1. Error Analysis

We manually analyzed the 78 questions in which the majority of the workers provided incorrect answers to yield some insight into why the workers had difficulty and to determine if accuracy could be improved further. In general, the majority of questions for which two or three users picked the wrong answer were difficult cases where the attachment point was to a verb or an adjective; when the attachment point was to a noun, users were almost always able to correctly identify it. For example, consider the first example from Table 3. The correct attachment to the PP is the verb "is", but all three users incorrectly selected the noun phrase "striking finding" instead.

Another issue that arose was when the sentence had two possible correct attachments, as shown in the second example in Table 3: "obtain regulatory approval" and "complete the transaction" are equally valid attachment points, but only "complete the transaction" was selected as the correct answer by our system. Therefore all users who chose "obtain regulatory approval" were marked as incorrect.

5.2. Worker Analysis

Figure 2 displays the questions in terms of the number of workers per question that responded correctly. The cases in which 2/3 and 3/3 workers responded correctly contribute to the accuracy results listed previously.

In Figure 3, we observe that agreement between workers is a strong indicator of accuracy and that unanimous agreement is more likely to be accurate (97%) than simple majority agreement (82%). In cases where there was no agreement between workers, the answer was always wrong; in a real-world annotation scenario, these cases could be discarded or resubmitted to Mechanical Turk in an attempt to obtain agreement. Finally, while these results indicate that three workers per question is sufficient for finding the correct answer in this domain, the 15% relative improvement in accuracy seen when moving from a plurality of 2 workers to 3 workers suggests that using more workers with a higher threshold for agreement might yield still stronger results.

6. Conclusion & Future Work

We have described a semi-automated approach to building a corpus of PP attachments using untrained annotators at low cost. With three workers, we obtain an accuracy of 92.2% using majority agreement. Furthermore, our analysis shows that accuracy increases when all three workers agree, suggesting that increasing the number of workers would further increase accuracy. Our error analysis shows that workers tend to agree when a PP modifies a noun, but make more mistakes when the PP modifies a VP or adjective.

We plan to extend this work by doing similar analysis on genres such as weblogs, which are not well represented in annotation projects. In order to replace the use of gold-standard parses from the Penn Treebank used in this article, we are creating a domain-independent system for extracting a valid set of PP-attachment options from natural language text and minimizing the options for disambiguation that are

provided to humans. In conjunction with online annotation marketplaces such as Mechanical Turk, this system would provide a methodology to obtain large quantities of valuable training data for statistical tools.

7. Acknowledgements

The authors would like to thank Kevin Lerman for his help in formulating the original ideas for this work. This material is based on research supported in part by the U.S. National Science Foundation (NSF) under IIS-05-34871 Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

8. References

Eneko Agirre, Timothy Baldwin, and David Martinez. 2008. Improving parsing and PP attachment performance with sense information. In *Proceedings of ACL*, pages 317–325, June.

Michaela Atterer and Hinrich Schütze. 2007. Prepositional phrase attachment without oracles. *Computational Linguistics*, 33(4):469–476.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of ACL*, pages 132–139.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637.

Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. 2004. Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 491–501.

Liangjie Hong and Brian D. Davison. 2009. A classification-based approach to question answering in discussion boards. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 171–178.

Gilly Leshed and Joseph 'Jofish' Kaye. 2006. Understanding how bloggers feel: recognizing affect in blog posts. In *CHI '06 extended abstracts on Human factors in computing systems*, pages 1019–1024.

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

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The nombank project: An interim report. In Adam Meyers, editor, *Proceedings of the HLT-NAACL Workshop on Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Sameer S. Pradhan and Nianwen Xue. 2009. Ontonotes: the 90% solution. In *Proceedings of NAACL*, pages 11–12.

Adwait Ratnaparkhi, Jeff Reynar, and Salim Roukos. 1994. A maximum entropy model for prepositional phrase attachment. In *Proceedings of HLT*, pages 250–255.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*, pages 254–263.

Jiri Stetina and Makoto Nagao. 1997. Corpus based pp attachment ambiguity resolution with a semantic dictionary. In *Proceedings of the Workshop on Very Large Corpora*, pages 66–80.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task: chunking. In *Proceedings of the CoNLL workshop on Learning Language in Logic*, pages 127–132.

Alexander S. Yeh and Marc B. Vilain. 1998. Some properties of preposition and subordinate conjunction attachments. In *Proceedings of COLING*, pages 1436–1442.

Jakub Zavrel, Walter Daelemans, and Jorn Veenstra. 1997. Resolving PP attachment ambiguities with memory-based learning. In *Proc. of the Workshop on Computational Language Learning (CoNLL'97), ACL*, pages 136–144.