PARSING AND ITS APPLICATIONS FOR CONVERSATIONAL SPEECH

Matthew Lease, Eugene Charniak, and Mark Johnson

Brown Laboratory for Linguistic Information Processing
Brown University
{mlease,ec,mj}@cs.brown.edu

ABSTRACT

This paper provides an introduction to recent work in statistical parsing and its applications for conversational speech, with particular emphasis on the relationship between parsing and detecting speech repairs. While historically parsing and repair detection have been studied independently, we present a line of research which has spanned the boundary between the two and demonstrated the efficacy of this synergistic approach. Our presentation highlights successes to date, remaining challenges, and promising future work.

1. INTRODUCTION

In the last decade, statistical methods for syntactic parsing have matured to the point where broad coverage, highly accurate, and efficient parsing of text is now a reality. The significance of this is that one can now reliably obtain structural information underlying language usage that can be exploited to create more accurate models of language production and provide insights for its interpretation. In recent years, this improved access to syntactic information has directly led to improved standards of performance in language modeling [1], speech recognition [2], disfluency detection [3], and machine translation [4].

In a parallel and largely separate track of research, speech recognition accuracy has also improved dramatically over the past 10 years. However, this work has largely restricted itself to word token recognition, ignoring "metadata" issues such as sentence boundary and disfluency detection. While automatic recognition of these phenomena would clearly benefit tasks like transcript cleanup, of greater significance is the need for such detection as a preprocessing step for downstream applications like machine translation and information extraction. While most existing work in metadata detection has focused on rich acoustic analysis [5], recent work in syntax-driven techniques has shown this latter approach to be equally effective, as well as demonstrating the enormous potential that exists for developing new synergies between these previously disparate approaches [6].

This work was supported by NSF grants LIS 9720368 and IIS0095940.

In this paper, we describe recent work in statistical parsing and its applications for conversational speech, emphasizing the relationship between parsing and detecting speech repairs. In Section 2, we introduce parsing and parser-based language modeling. Section 3 highlights challenges in parsing speech with particular regard to speech repairs. In Section 4, we describe repairs in more detail and present an approach for detecting them using our parser-based language model. Finally, Section 5 concludes and discusses important issues for future work.

2. PARSING AND LANGUAGE MODELING

The goal of syntax is to find a systematic set of rules (a grammar) that accurately models the infinite number of ways words can (and cannot) be combined to form meaningful phrases, and how such phrases can further combine to create meaningful sentences. While there is no English grammar today that is universally accepted by linguists, the grammar induced by the Penn Treebank (PTB) [7] has been used to improve the state-of-the-art on a number of tasks [1, 2, 3, 4].

The task of parsing can be defined most simply as finding one or more syntactic analyses of a given sentence that are consistent with a particular grammar. To give an example, consider the sentence "John saw the man with the binoculars." Figure 1 shows two possible structures for this sentence. Note that the two correspond to different meanings: in Fig. 1(a), the man has the binoculars, while in Fig. 1(b) John is using the binoculars to see the man. This simple example highlights a couple of interesting issues. On one hand, we see how syntax and semantics intertwine: finding the correct syntactic analysis for a given sentence can help shed light on its intended meaning. At the same time, the example also introduces the problem of syntactic ambiguity that is rampant in practice: a typical length sentence will have hundreds of competing syntactic analyses from which the parser must select one as most likely.

Our state-of-the-art PTB-based parser [8] achieves approximately 90% labelled precision and recall in matching annotator analyses, as measured on PTB's Wall Street Journal corpus, the traditional benchmark of the statisti-

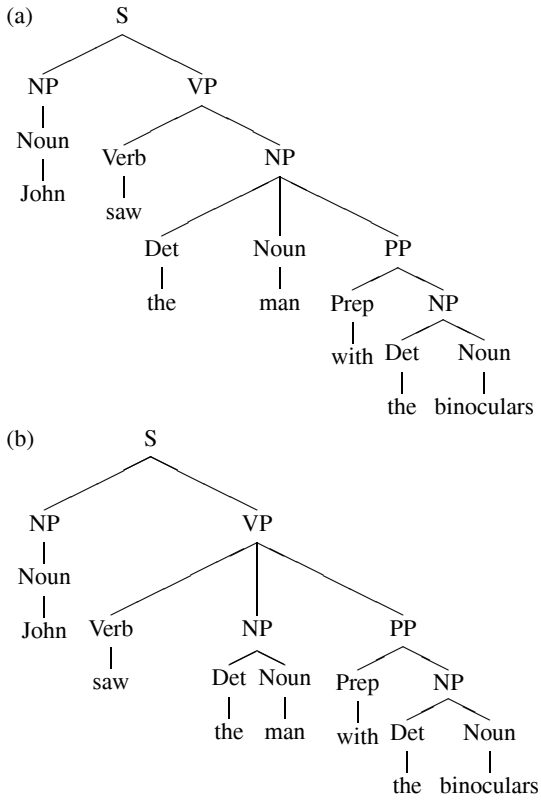

Fig. 1. Two structures for an ambiguous sentence.

cal parsing community. The parser is formulated as a lexicalized probabilistic context-free grammar (PCFG), with probabilities of the various syntactic rules seen in the treebank estimated via maximum likelihood. The use of lexicalized conditioning should be emphasized. Whereas a traditional, non-lexicalized CFG would produce the same analysis whether the last word of our example were "binoculars" or "hat", a lexicalized CFG can learn (from treebank examples) to prefer Figure 1(b)'s analysis in the case of "binoculars" and Figure 1(a)'s analysis in the case of "hat".

A parse π's probability is determined by a top-down process of guessing each constituent c's pre-terminal tag t, then its lexical head h (its most important word syntactically), and finally its expansion e into other constituents, given its label l (e.g. is it a noun or verb phrase) and relevant history H (information outside c that our probability model deems important). Thus we have

$$p(\pi) = \prod_{c \in \pi} p(t \mid l, H) \cdot p(h \mid t, l, H) \cdot p(e \mid l, t, h, H)$$

To find the most likely parse for a given sentence s, one just selects the π that maximizes the conditional probability. This parsing model is of further interest because its generative model immediately leads to a way to perform syntax-based language modeling: to estimate the probability of s, just sum over a significant sample of its possible parses. In comparison to the ubiquitous trigram, the relative efficacy of such syntax-based language modeling is clear, as demonstrated by the perplexity and WER results in Table 1.

	Perplexity		
Model	**Alone**	**+Trigram**	**WER**
Trigram	≈ 167	–	13.7
Xu [2]	151.2	144.2	12.3
Roark [9]	152.3	137.3	12.7
Charniak [1]	130.2	126.1	11.9

Table 1. Perplexity results of syntax-based language models on a "speech-like" version of WSJ. Trigram interpolation is with constant 0.36 for all models. Word Error Rate (WER) is for n–best list rescoring on HUB–1 lattices [10].

3. PARSING SPEECH

While statistical parsing of textual corpora has been studied for more than a decade, new challenges arise when one applies parsing to conversational speech. Not only are clearly textual features such as punctuation and capitalization absent, but we even lack clear boundaries as to where one "sentence-like unit" (SU) ends and the next begins, especially as speakers interrupt one another and provide backchannel feedback. Most spontaneous speech also abounds with disfluencies such as partial words, filled pauses (e.g., "uh", "um"), explicit editing terms (e.g., "I mean"), and parenthetical asides.

One type of disfluency that has proven particularly problematic for parsing is speech repairs: when a speaker amends what he is saying mid-sentence (see Figure 2). Following the analysis of [12], a speech repair can be understood as consisting of three parts: the *reparandum* (the material repaired), the *editing phrase* (that is typically either empty or consists of a filler), and the *repair*. Speech repairs are difficult to model in HMM or PCFG models (e.g. the PCFG model described in Section 2) because these models can induce only linear or tree-structured dependencies between words. However, the relationship between reparandum and repair seems to be quite different: the repair is often a "rough copy" of the reparandum, using the same or very similar words in roughly the same order [13]. In other words, a speech repair seems to involve "crossed" dependencies between the reparandum and the repair, as Figure 2 shows.

Such crossed-dependencies tend to cause collateral damage to the entire syntactic analysis produced by a PCFG. For this reason, we have adopted the approach of first detecting and filtering out repairs, and then parsing the remainder of the SU. Such filtering is reasonable since repairs only arise from production error, and thus may be safely removed

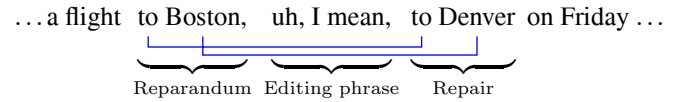

Fig. 2. The structure of a typical repair, with crossing dependencies between reparandum and repair.

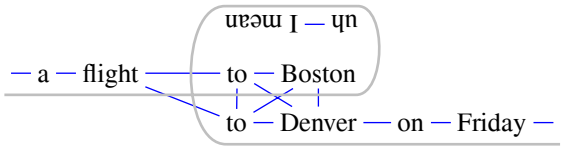

Fig. 3. The "helical" dependency structure induced by the generative model of speech repairs corresponding to Figure 2.

without any impact on meaning. Using this approach, we have achieved parsing precision and recall of about 86% on the Switchboard corpus of telephone conversations (assuming known SU boundaries) [11]. With oracle knowledge of repairs, this accuracy improves to 88%, so parsing stands to benefit as repair detection techniques improve.

4. DETECTING SPEECH REPAIRS

Unlike HMMs and PCFGs, mildly context-sensitive grammars such as Tree Adjoining Grammars (TAGs) are capable of describing crossing dependencies such as exhibited by speech repairs. To effectively model both the crossed-dependencies of speech repairs and the more usual linear or tree-structured dependencies of non-repaired speech, we have applied the noisy channel paradigm [3]. We begin by imagining that speakers intend to say source sentences S (with no repairs), but may mistakenly insert one or more repairs, producing observed sentences O. Our goal, then, is for each observed sentence to recover the most likely source sentence $\widehat{s}$. Applying Bayes Rule, we can formulate this problem in canonical noisy channel form

$$\widehat{s} \;=\; \operatorname*{argmax}_S \mathrm{P}(S|O) = \operatorname*{argmax}_S \mathrm{P}(O|S)\mathrm{P}(S)$$

The *channel model* $\mathrm{P}(O|S)$ defines a stochastic mapping of source sentences into observed sentences via the optional insertion of one or more repairs. Similarly, our *language model* $\mathrm{P}(S)$ defines a probability distribution over source sentences. This is the same general setup that is used in statistical speech recognition and machine translation, and in these applications syntax-based language models yield state-of-the-art performance [4, 10]. In this case, the channel model is realized as a TAG, and we train our parser-based language model (Section 2) on sentences with the speech repairs removed. Figure 3 shows the combined model's dependency structure for the repair of Figure 2. If we trace the temporal word string through this dependency structure, aligning words next to the words they are dependent on, we obtain a "helical" type of structure familiar from genome models, and in fact TAGs are being used to model genomes for very similar reasons.

The output of our noisy channel model is a set of candidate repairs, which we then rescore using a maximum-entropy model [6]. Using the MaxEnt model makes it relatively easy to experiment with a wider range of features beyond the TAG and PCFG probabilities. For example, we added features based on the local context of the reparandum that were used in an earlier algorithm [11]. An informal error analysis of the two repair detection algorithms [3, 11] suggested that the noisy channel model was better at detecting moderately long repairs, but the earlier classifier was better at detecting very short repairs. Also, as parses let us identify the syntactic context in which each speech repair occurs, we have found the category labels immediately dominating, preceding, and following the repair to be useful features. Finally, we have incorporated features based on prosodic information: word-by-word interruption point (IP) probabilities produced by ICSI-SRI-UW [14].

In the recent Fall 2004 Rich Transcription blind evaluation, metadata extraction systems competed in three unique disfluency detection tasks. *Edit Word Detection* requires determining which words were part of the reparandum region of a speech repair. *Filler Word Detection* involves identifying both which words are part of a filler phrase and the type of filler (e.g. filled pause vs. discourse marker). *Interruption Point Detection* requires detecting the point at which speech became disfluent. For all three tasks, the goal is to minimize a simple error metric: the number of mistakes (false positives + false negatives) divided by the actual number of events (true positives). Performance was measured on two types of input: manually annotated (reference) words and automatically recognized speech-to-text

Detection Task	STT	Reference
Edit Words	76.25	46.08
Filler Words	39.93	23.69
Interruption Points	55.88	28.60

Table 2. Error rates on several metadata extraction tasks

(STT) tokens. For all three tasks, and on both types of input, our noisy-channel, MaxEnt rescoring model was the top performer in the evaluation [6]. Official results are given in Table 2. Filler word detection, substantially easier than edit word detection, was achieved via a few hand-crafted, deterministic rules. Our interruption point predictions were explicitly determined by predicted edit and filler words.

5. CONCLUSION FUTURE WORK

We have reviewed recent work in statistical parsing and its applications for conversational speech, with particular emphasis on the relationship between parsing and detecting speech repairs. We now expand upon some of challenges presented by conversational speech (Section 3) and identify some other interesting areas for future work.

How can prosodic cues be most effectively leveraged in parsing? To what degree can syntax be leveraged to support SU boundary detection? Can parsing be made more robust to ASR and SU boundary detection errors? How should parsing accuracy be measured in the presence of these errors? A forthcoming Johns Hopkins workshop led by Mary Harper will investigate these and related issues.

Little has been done to date in using partially labeled and unlabeled training to improve syntax-based language models. In particular, there is a common misperception that these models require hand-parsed training data, limiting the amount of data available to them. This is not the case. We have found that unlabeled data can be used to improve perplexity and word-error rates [10], and there is a significant opportunity for further error reductions via this avenue.

Finally, there has also been encouraging work in working directly off of the word lattices produced by speech recognition systems [10]. By exploiting redundancy across different paths through the lattice, such techniques have the opportunity to dramatically reduce complexity in computing alternatives. Furthermore, working off the lattice provides significantly more discrimination power, as well as a wider range of candidates, than is possible with the typical n-best rescoring approach.

6. REFERENCES

[1] E. Charniak, "Immediate-head parsing for language models," in *39th Annual Meeting of the ACL*, 2001.

[2] P. Xu, C. Chelba, and F. Jelinek, "A study on richer syntactic dependencies for structured language modeling," in *40th Annual Meeting of the ACL*, 2002.

[3] M. Johnson and E. Charniak, "A TAG-based noisy channel model of speech repairs," in *Proc. of the 42nd Annual Meeting of ACL*, 2004, pp. 33–39.

[4] E. Charniak, K. Knight, and K. Yamada, "Syntax-based language models for statistical machine translation," in *Intl. Machine Translation Summit*, 2003.

[5] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper, "Structural metadata research in the ears program," in *2005 IEEE Intl. Conference on Acoustics, Speech, and Signal Processing*, 2005.

[6] M. Johnson, E. Charniak, and M. Lease, "An improved model for recognizing disfluencies in conversational speech," in *Proc. of the Rich Transcription 2004 Fall Workshop*, 2004.

[7] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[8] E. Charniak, "A maximum-entropy-inspired parser," in *1st Annual Meeting of the NAACL*, 2000.

[9] B. Roark, "Probabilistic top-down parsing and language modeling," *Computational Linguistics*, vol. 27, no. 2, pp. 249–276, 2001.

[10] K. Hall and M. Johnson, "Language modeling using efficient best-first bottom-up parsing," in *Automatic Speech Recognition and Understanding Workshop*, 2003.

[11] E. Charniak and M. Johnson, "Edit detection and parsing for transcribed speech," in *Proc. of the 2nd Annual Meeting of the NAACL*, 2001, pp. 118–126.

[12] E. Shriberg, *Preliminaries to a Theory of Speech Disfluencies*, Ph.D. thesis, University of California, Berkeley, 1994.

[13] P. A. Heeman and J. F. Allen, "Speech repairs, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue," *Comput. Linguist.*, vol. 25, no. 4, pp. 527–571, 1999.

[14] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, B. Peskin, and M. Harper, "The ICSI-SRI-UW metadata extraction system," in *Proc. of the Intl. Conference on Spoken Language Processing*, 2004.