Proceedings of the Conference on Empirical Methods in Natural
Language Processing (EMNLP), Philadelphia, July 2002, pp. 49-54.
Association for Computational Linguistics.

Parsing and Disfluency Placement

Donald Engel[†] and **Eugene Charniak**[‡] and **Mark Johnson**[‡]
Department of Physics, University of Pennsylvania[†]
Brown Laboratory for Linguistic Information Processing[‡]
Brown University

Abstract

It has been suggested that some forms of speech disfluencies, most notable interjections and parentheticals, tend to occur disproportionally at major clause boundaries [6] and thus might serve to aid parsers in establishing these boundaries. We have tested a current statistical parser [1] on Switchboard text with and without interjections and parentheticals and found that the parser performed better when not faced with these extra phenomena. This suggest that for current parsers, at least, interjection and parenthetical placement does not help in the parsing process.

1 Introduction

It is generally recognized that punctuation helps in parsing text. For example, Roark [5] finds that removing punctuation decreases his parser's accuracy from 86.6% to 83.8%. Our experiments with the parser described in [1] show a similar falloff. Unfortunately spoken English does not come with punctuation, and even when transcriptions add punctuation, as in the Switchboard [4] corpus of transcribed (and parsed) telephone calls, it's utility is small [5] For this and other reasons there is considerable interest in finding other aspects of speech that might serve as a replacement.

One suggestion in this vein is that the placement of some forms of speech errors might encode useful linguistic information. Speech, of course, contains many kinds of errors that can make it more difficult to parse than text. Roughly speaking the previously mentioned Switchboard corpus distinguishes three kinds of errors:

- interjections (filled pauses) — "I, um, want to leave"
- parentheticals — "I, you know, want to leave"
- speech repairs — "I can, I want to leave"

Of these, speech repairs are the most injurious to parsing. Furthermore, even if one's parser can parse the sentence as it stands, that is not sufficient. For example, in "I can, I want to leave", it is not necessarily the case that the speaker believes that he or she can, in fact, leave, only that he or she wants to leave. Thus in [2] speech repairs were first detected in a separate module, and deleted before handing the remaining text to the parser. The parser then produced a parse of the text without the repaired section.

The other two kinds of errors, interjections, and parentheticals, (henceforth INTJs and PRNs) are less problematic. In particular, if they are left in the text either their semantic content is compatible with the rest of the utterance or there is no semantic content at all. For example, Table 1 gives the 40 most common INTJs, which comprise 97% of the total. (Unlisted INTJs comprise the remaining 3%.) They are easily recognized as not carrying much, if any, content.

PRNs are more diverse. Table 2 lists the 40 most common PRNs. They only comprise 65% of all cases, and many do contain semantics content. In such cases, however, the semantic content is compatible with the rest of the sentence, so leaving them in is perfectly acceptable. Thus [2], while endeavoring to detect and remove speech repairs, left interjections and parentheticals in the text for the parser to cope with.

Indeed [6] finds that both interjections and parentheticals tend to occur at major sentence boundaries. Also [7] suggest that this prop-

Phrase	Num. of INTJs	Percent
uh	17609	27.44
yeah	11310	17.62
uh-huh	7687	11.97
well	5287	8.238
um	3563	5.552
oh	2935	4.573
right	2873	4.477
like	1772	2.761
no	1246	1.941
okay	1237	1.927
yes	982	1.530
so	651	1.014
oh yeah	638	0.994
huh	558	0.869
now	410	0.638
really	279	0.434
sure	276	0.430
oh okay	269	0.419
see	261	0.406
oh really	260	0.405
huh-uh	185	0.288
wow	174	0.271
bye-bye	174	0.271
exactly	156	0.243
all right	146	0.227
yep	115	0.179
boy	111	0.172
oh no	102	0.158
bye	98	0.152
well yeah	91	0.141
gosh	91	0.141
oh gosh	88	0.137
oh yes	84	0.130
hey	75	0.116
uh yeah	71	0.110
anyway	71	0.110
oh uh-huh	70	0.109
say	63	0.098
oh goodness	61	0.095
uh no	56	0.087

Table 1: The 40 Most Common Interjections

Phrase	Num. of PRNs	Percent
you know	431	37.02
I mean	105	9.020
I think	86	7.388
I guess	67	5.756
You know	44	3.780
I don’t know	38	3.264
let’s see	11	0.945
I I mean	10	0.859
I ’d say	9	0.773
I ’m sure	7	0.601
excuse me	6	0.515
what is it	6	0.515
I would say	5	0.429
you you know	5	0.429
let ’s say	5	0.429
I think it ’s	4	0.343
I ’m sorry	4	0.343
so to speak	3	0.257
I guess it ’s	3	0.257
I don’t think	3	0.257
I think it was	3	0.257
I would think	3	0.257
it seems	3	0.257
I guess it was	2	0.171
I know	2	0.171
I I I mean	2	0.171
seems like	2	0.171
Shall we say	2	0.171
I guess you could say	2	0.171
You’re right	2	0.171
I believe	2	0.171
I think it was uh	2	0.171
I say	2	0.171
What I call	2	0.171
I don’t know what part of New Jersey you’re in but	2	0.171
I should say	2	0.171
I guess not a sore thumb	1	0.085
I ’m trying to think	1	0.085
And it’s hard to drag her away	1	0.085
I don’t know what you call that	1	0.085

Table 2: The 40 Most Common Parentheticals

erty accounts for their observation that removing these disfluencies does not help in language modeling perplexity results. This strongly suggests that INTJ/PRN location information in speech text might in fact, improve parsing performance by helping the parser locate constituent boundaries with high accuracy. That is, a statistic parser such as [1] or [3] when trained on parsed Switchboard text with these phenomena left in, might learn the statistical correlations between them and phrase boundaries just as they are obviously learning the correlations between punctuation and phrase boundaries in written text.

In this paper then we wish to determine if the presence of INTJs and PRNs do help parsing, at least for one state-of-the-art statistical parser [1].

2 Experimental Design

The experimental design used was more complicated than we initially expected. We had anticipated that the experiments would be conducted analogously to the "no punctuation" experiments previously mentioned. In those experiments one removes punctuation from all of the corpus sentences, both for testing and training, and then one reports the results before and after this removal. (Note that one must remove punctuation from the training data as well so that it looks like the non-punctuated testing data it receives.) Parsing accuracy was measured in the usual way, using labeled precision recall. Note, however, and this is a critical point, that precision and recall are only measured on non-preterminal constituents. That is, if we have a constituent

```
(PP (IN of)
    (NP (DT the) (NN book)))
```

our measurements would note if we correctly found the PP and the NP, but not the preterminals IN, DT, and NN. The logic of this is to avoid confusing parsing results with part-of-speech tagging, a much simpler problem.

Initially we conducted similarly designed experiments, except rather than removing punctuation, we removed INTJs and PRNs and compared before and after precision/recall numbers. These numbers seemed to confirm the anticipated results: the "after" numbers, the numbers without INTJ/PRNs were significantly worse, suggesting that the presence of INTJ/PRNs helped the parser.

Unfortunately, although fine for punctuation, this experimental design is not sufficient for measuring the effects of INTJ/PRNs on parsing. The difference is that punctuation itself is not measured in the precision-recall numbers. That is, if we had a phrase like

```
(NP (NP (DT a) (NN sentence))
    (, ,)
    (ADJP (JJ like)
          (NP (DT this) (DT one))))
```

we would measure our accuracy on the three NP's and the ADJP, but not on the preterminals, and it is only at the preterminal level that punctuation appears.

The same cannot be said for INTJ/PRNs. Consider the (slightly simplified) Switchboard parse for a sentence like "I, you know, want to leave":

```
(S (NP I)
   (PRN , you know ,)
   (VP want (S to leave)))
```

The parenthetical PRN is a full non-terminal and thus is counted in precision/recall measurements. Thus removing preterminals is changing what we wish to measure. In particular, when our initial results showed that removal of INTJ/PRNs lowered precision/recall we worried that it might be that INTJ/PRNs are particularly easy to parse, and thus removing them made things worse, not because of collateral damage on our ability to parse other constituents, but simply because we removed a body of easily parseable constituents, leaving the more difficult constituents to be measured. The above tables of INTJs and PRNs lends credence to this concern.

Thus in the experiments below all measurements are obtained in the following fashion:

1. The parser is trained on switchboard data with/without INTJ/PRNs or punctuation, creating eight configurations: 4 for neither, both, just INTJs, and just PRNs, times two for with and without punctuation. We tested with and without punctuation to confirm Roark's earlier results showing that

they have little influence in Switchboard text.

2. The parser reads the gold standard testing examples and depending on the configuration INTJs and/or PRNS are removed from the gold standard parse.

3. Finally the resulting parse is compared with the gold standard. However, any remaining PRNs or INTJs are ignored when computing the precision and recall statistics for the parse.

To expand a bit on point (3) above, for an experiment where we are parsing with INTJs, but not PRNs, the resulting parse will, of course, contain INTJs, but (a) they are not counted as present in the gold standard (so we do not affect recall statistics), and (b) they are not evaluated in the guessed parse (so if one were labeled, say, an S, it would not be counted against the parse). The intent, again, is to not allow the results to be influenced by the fact that interjections and parentheticals are much easier to find than most (if not) all other kinds of constituents.

3 Experimental Results

As in [2] the Switchboard parsed/merged corpus directories two and three were used for training. In directory four, files sw4004.mrg to sw4153.mrg were used for testing, and sw4519.mrg to sw4936 for development. To avoid confounding the results with problems of edit detection, all edited nodes were deleted from the gold standard parses.

The results of the experiment are given in table 3. We have shown results separately with and without punctuation. A quick look at the data indicates that both sets show the same trends but with punctuation helping performance by about 1.0% absolute in both precision and recall. Within both groups, as is always the case, we see that the parser does better when restricted to shorter sentences (40 words and punctuation or less). We see that removing PRNs or INTJs separately both improve parsing accuracy (e.g., from 87.201% to 87.845—that the effect of removing both is approximately additive (e.g., from 87201% to 88.863%, again on the with-punctuation data). Both with and without punctuation results hint that removing parentheticals was usually more helpful than removing interjections. However in one case the reverse was true (with-punctuation, sentences ≤ 40) and in all cases the differences are at or under the edge of statistical reliability. In contrast, the differences between removing neither, removing one, or removing both INJs and PRNs are quite comfortably statistically reliable.

Punc.	PRN	INTJ	Sentences	Sentences
			≤ 40	≤ 100
+	+	+	88.93	87.20
+	+	-	89.44	87.85
+	-	+	89.13	87.99
+	-	-	90.00	88.86
-	+	+	87.40	86.23
-	+	-	88.0	86.8
-	-	+	88.41	87.45
-	-	-	89.13	88.30

Table 3: Average of labeled precision/recall data for parsing with/without parentheticals/interjections

4 Discussion

Based upon Tabel 3 our tentative conclusion is that the information present in parentheticals and interjections does not help parsing. There are, however, reasons that this is a *tentative* conclusion.

First, in our effort to prevent the ease of recognizing these constructions from giving an unfair advantage to the parser when they are present, it could be argued that we have given the parser an unfair advantage when they are absent. After all, even if these constructions are easily recognized, the parser is not perfect on them. While our labeled precision/recall measurements are made in such a way that a mistake in the label of, say, an interjection, would not effect the results, a mistake on it's position typically *would* have an effect because the positions of constituents either before or after it would be made incorrect. Thus the parser has a harder task set for it when these constituents are left in.

It would be preferable to have an experimental design that would somehow equalize things, but we have been unable to find one. Furthermore it is instructive to contrast this situation with that of punctuation in Wall Street

Journal text. If we had found that parsing without punctuation made things easier a similar argument could be made that the without-punctuation case was given an unfair advantage since it had a lot fewer things to worry about. But punctuation in well-edited text contains more than enough information to overcome the disadvantage. This does not seem to be the case with INTJs and PRNs. Here the net information content here seems to be negative.

A second, and in our estimation more serious, objection to our conclusion is that we have only done the experiment with one parser. Perhaps there is something specific to this parser that systematically underestimates the usefulness of INTJ/PRN information. While we feel reasonably confident that any other current parser would find similar effects, it is at least possible to imagine that quite different parsers might not. Statistical parsers condition the probability of a constituent on the types of neighboring constituents. Interjections and parentheticals have the effect of increasing the kinds of neighbors one might have, thus splitting the data and making it less reliable. The same is true for punctuation, of course, but it seems plausible that well edited punctuation is sufficiently regular that this problem is not too bad, while spontaneous interjections and parentheticals would not be so regular. Of course, finding a parser design that might overcome this problem (assuming that this *is* the problem) is far from obvious.

5 Conclusion

We have tested a current statistical parser [1] on Switchboard text with and without interjections and parentheticals and found that the parser performs better when not faced with these extra phenomena. This suggest that for current parsers, at least, interjection and parenthetical placement does not help in the parsing process.

This is, of course, a disappointing result. The phenomena are not going to go away, and what this means is that there is probably no silver lining.

We should also note that the idea that they might help parsing grew from the observation that interjections and parentheticals typically occur at major clause boundaries. One might then ask if our results cast any doubt on this claim as well. We do not think so. Interjections and parentheticals do tend to identify clause boundaries. The problem is that many other things do so as well, most notably normal grammatical word ordering. The question is whether the information content of disfluency placement is sufficient to overcome the disruption of word ordering that it entails. The answer, for current parsers at least, seems to be ”no”.

6 Acknowledgements

We would like to acknowledge the members of the Brown Laboratory for Linguistic Information Processing, This research has been supported in part by NSF grants IIS 0085940, IIS 0112435, and DGE 9870676.

References

1. Charniak, E. *A maximum-entropy-inspired parser*. In *Proceedings of the 2000 Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, New Brunswick NJ, 2000, 132–139.

2. Charniak, E. and Johnson, M. *Edit Detection and Parsing for Transcribed Speech*. In *Proceedings of the North American Assocation for Computational Linguistics 2001*. 2001, 118–126.

3. Collins, M. J. *Three generative lexicalized models for statistical parsing*. In *Proceedings of the 35th Annual Meeting of the ACL*. 1997, 16–23.

4. Godfrey, J. J., Holliman, E. C. and McDaniel, J. *SWITCHBOARD: Telephone speech corpus for research and development*. . In *Proceedings IEEE Conference on Acoustics, Speech and Signal Processing*. San Francisco, 1992, 517–520 .

5. Roark, B. *Robust Probabilistic Predictive Syntactic Processing: Motivations, Models, and Applications*. In *Ph.D. thesis*. Department of Cognitive Science, Brown University, Providence, RI, 2001.

6. Shriberg, E. E. *Preliminaries to a Theory of Speech Disfluencies*. In *Ph.D. Dissertation*. Department of Psychology, University of California-Berkeley, 1994.

7. Stolcke, A. and Shriberg, E. *Automatic linguistic segmentation of conversational speech.* In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP-96)*. 1996.