The Utility of Filled Pauses, Interjections, and Parentheticals in Parsing Conversational Language

by

Donald Edward Engel

B.S., Brown University, 2000

Thesis

Submitted in partial fulfillment of the requirements for the Degree of

Master of Science in the Department of Computer Science at Brown University

May 2001

ABSTRACT

It has been hypothesized [5,7] that the information contained by the placement and content of filled pauses, interjections, and parentheticals is useful in the parsing of conversational language. To investigate the impact of the inclusion of such information on the performance of a parser, we have built wrappers to work between the "Switchboard" corpus [3] and an existing parser [1]. These wrappers each removed one or more types of linguistic elements from the corpus. Training and testing on each wrapped corpus, we achieve different success rates. Our data indicates that some forms of misspeaking have little effect, but others impede the functionality of the parser.

1. INTRODUCTION

1.1 Motivation and Overview

While there has been some very good work in the processing of transcribed speech, the existing parsers and techniques are largely derivative of techniques developed for application to printed media. In spoken speech, particularly in improvised, conversational speech, several additional features exist. Disfluencies are often part of transcribed corpora, yet the role this information can play in linguistic analysis is largely unexplored. Errors in speaking occur in several varieties, and it is not clear that all of these errors are of similar content or utility. In [5] and [7], there are three categories of disfluency type established, defined by what must be done to correct for them. They are:

Filled Pauses. Utterances habitually inserted during a pause (e.g. "um" or "uh").

That's uh interesting.

→ That's interesting

Repetitions. Repetitions of words that are (without the repetition) intended.

That's that's interesting

→ That's interesting.

Deletions. Words that have no correlation to the surrounding sequence.

You that's interesting.

→ That's interesting.

While not all errors are of these forms, we know from [9] that they account for 85% of the disfluencies in the source we used for transcribed speech [3]. In this paper, we explore the relative utility of several forms of misspeaking, hoping to improve the functionality of speech processing by knowing which errors to heed and which to disregard.

1.2 Existing Tools and Resources Utilized

In any parsing experiments, a corpus of text is used for testing purposes, and often for training purposes as well. In order to be certain that the statistics are generally applicable, the portion of a corpus used for training will be distinct from the section used for testing. In our experiment, we use the Switchboard corpus [3], which consists of transcribed telephone conversations tagged with part-of-speech information and trees that represent the sentence (phrasal) structure. Crucial to our experiment, the corpus contains what was said, and not a cleaned-up version of what was meant. There are therefore many instances of disfluency (more details in section 1.3 below). Although it was not explicitly spoken, implied punctuation is included in the corpus.

A traditional statistical parsing experiment works in three phases. First, a program collects statistics on a designated training portion of a corpus, taking advantage of tag and tree information provided therein. Second, a parser attempts to generate trees for a testing portion of the corpus, ignoring the trees already therein. In the final stage, the performance of the parser is evaluated by comparing the actual trees in the testing portion with those generated in the second phase. Rather than develop a new parser for the purpose of our experiment, we use the parser originally developed by Charniak [1] to work on ideal text, and later modified by him to handle transcribed speech [2].

1.3 The Forms and Prevalence of Misspeaking

Within the Switchboard corpus, there are four tags that correspond to errors in speech. An UH is an inserted word such as "yeah," "um," and "sure." An interjection (tagged as INTJ) is usually a phrase containing one or more words tagged as UH. A parenthetical is an otherwise legitimate phrase that is inserted into a sentence typically without adding meaning, such as "you know," or "I mean." Phrases tagged as EDITED are places where a speaker disregarded that portion of their speech and began again. For example, "we, we hire these wonderful people" and "you, I try to really watch it." Recalling section 1.1 above, an UH tags and INTJ phrases correspond roughly with a filled pauses, while an EDITED phrase corresponds with deletions, repetitions, and uncategorized errors.

Table 1. Sample Disfluencies

UH	(UH Uh-huh) (UH Uh) (UH like)
PRN	(PRN (, ,) (S (NP-SBJ (PRP you)) (VP (VBP know))))) (PRN (S (NP-SBJ (PRP I)) (VP (VBP mean))))

INTJ	((INTJ (INTJ (UH Okay)) (, ,) (INTJ (UH um)) (, ,) (-DFL- E_S))) (INTJ (UH like))
EDITED	…(EDITED (RM (-DFL- \[)) (S (NP-SBJ (PRP it)) (VP-UNF (VBD was))) (, ,) (IP (-DFL- \+))) (NP-SBJ (PRP it)) (VP (VBD was) …

In the Table 1 above, the additional tags have the following properties

Table 2. Some Switchboard Tags

E_S	End-of-sentence marker.
IP	The tag (IP +) marks the end of a restart. All restarts are in EDITED.
NP-SBJ	The noun phrase that is the subject of a sentence.
PRP	Personal Pronoun - I, me, you, him, etc.
RM	The tag (RM [) marks the beginning of a restart. All restarts are in EDITED.
S	A simple declarative sentence.
VBP	A verb that is non-third-person, singular, present tense
VBD	A past tense verb.
VP	A verb phrase. Modifications to VP can be added with a dash (see VP-UNF)
VP-UNF	An unfinished verb phrase. -UNF is added to any unfinished phrase.
-DFL-	The tagged item is not a word, but a code. See E_S, IP, and RM.

In the Switchboard corpus, interjections comprise 23.6% of the top-level structures (phrases that have no parent phrase), and words tagged as "UH" are 7.9% of the words spoken. If this information were not of value, discarding it would significantly reduce the sizes of the problems at hand. If it is of value, the prevalence suggests that this value may be worthy of consideration and optimization.

2. RESULTS AND ANALYSIS

2.1 Results from related work

The 2001 PhD thesis of Brian Roark [4] has already noted that punctuation is of less utility in transcribed speech than in professionally printed media. This difference is interesting to note, as it could be seen to indicate that spoken word orders contain more information required for estimating sentence boundaries than is contained by printed text.

On the other hand, it could mean that punctuation in the transcribed corpus is less consistent than punctuation in a printed journal. Either way, the difference is worthy of mention.

Stolcke and Shriberg [6] explore the hypothesis that "probability estimates for words after a disfluency are more accurate if conditioned on the intended fluent sequence." They find that this is true, but yields only "small improvements." In our approach, we explore the impact disfluencies more categorically, and categorize disfluencies by tree type (INTJ, PRN) rather than by the categories described in Section 1.1.

2.2 Our Results

The measure we have used to determine the quality of our parser's performance is the average of the labeled precision and labeled recall. Because the parser being utilized [2] already handles EDITED nodes in a specialized way, we focus our attention on the disfluencies UH, PRN, and INTJ. We have also experimented on the inclusion and exclusion of punctuation for two reasons. First, the punctuation was not spoken in the dialogues, but rather added by the transcribers. Punctuation is often, but not always, used to separate disfluencies from the rest of a sentence, and so the residual punctuation, were it not removed, might offer information which should no longer be present. Second, the punctuation serves a similar role in written text to what we might expect disfluencies to serve in speech. Namely, it occurs at regular phrasal boundaries. Any parallels that would become apparent between punctuation and disfluencies would therefore be of interest.

Due to confusion caused by removal of UH without removal of punctuation, these two were paired instead of including a case where only UH words were removed. The difficulty was caused by cases where there were one or more UH tags interspersed with punctuation, such as "uh, yeah, sure." On removal of the UH words, this would leave only ", ,". Leaving one or more pieces of punctuation in the corpus in the place of UH words was not desired, as this information was only available because the UH words had been in the corpus before it was modified by the wrapper. Such extraneous punctuation could have construed the same information to the parser as the UH words themselves.

Table 3. Overall Results

Punctuation	UH	PRN	INTJ	Sentences < 40	Sentences < 100
-	-	+	+	*87.802%*	*86.574%*
-	+	+	+	*88.052%*	*86.909%*
+	+	+	-	89.411%	87.802%
+	+	+	+	89.592%	87.956%
-	-	-	-	89.054%	88.186%
+	+	-	-	90.000%	88.863%
+	+	-	+	90.076%	88.902%

The above data is sorted by the average of the labeled precision and labeled recall of sentences with length under 100 words. It is interesting to note that the data for sentences of length less than forty words is not strictly increasing. This demonstrates that inclusion information of the same category may have different effects on sentences of differing lengths.

We also note that the punctuation, added by the transcribers, is of little use to the parser, which contradicts the results from corpora derived from printed media. Also from the data above, it seems removing parentheticals and interjections is more effective than removing punctuation and UHs.

3. DISCUSSION AND CONCLUSIONS

3.1 The Effect of UHs and Interjections

As explained in Section 2.2, UH tags were removed simultaneously with punctuation information. In so doing, the parser fared worse than it had under any other circumstances, which may suggest that UH information is useful to the parser. However, UH tags usually occur on only a small set of words (see Table 4), and the performance hit could be due to the credit a parser would normally receive for making a fairly easy assignments of the interjection tag INTJ to sets of these UH words. To avoid this ambiguity, alternative methods of evaluation were considered, but none seemed promising. A fairly straightforward approach would be to remove everything tagged as UH or INTJ from the corpus after guessing and before evaluation. However, this method could easily create a situation where words exist in the guessed corpus that were removed in the reference (gold standard) corpus, or vice versa. Such alignment errors are not allowed by the scoring metric applied [2].

Table 4. The 40 Most Common Interjections

Phrase	# of INTJs	% of INTJs	Phrase	# of INTJs	% of INTJs
uh	17609	27.44	huh-uh	185	0.288
yeah	11310	17.62	wow	174	0.271
uh-huh	7687	11.97	bye-bye	174	0.271
well	5287	8.238	exactly	156	0.243
um	3563	5.552	all right	146	0.227
oh	2935	4.573	yep	115	0.179
right	2873	4.477	boy	111	0.172
like	1772	2.761	oh no	102	0.158
no	1246	1.941	bye	98	0.152
okay	1237	1.927	well yeah	91	0.141
yes	982	1.530	gosh	91	0.141
so	651	1.014	oh gosh	88	0.137
oh yeah	638	0.994	oh yes	84	0.130

huh	558	0.869	hey	75	0.116
now	410	0.638	uh yeah	71	0.110
really	279	0.434	anyway	71	0.110
sure	276	0.430	oh uh-huh	70	0.109
oh okay	269	0.419	say	63	0.098
see	261	0.406	oh goodness	61	0.095
oh really	260	0.405	uh no	56	0.087

(unlisted interjections comprise the remaining 3% of interjections)

Without a better metric, it seems plausible that the UH information was of little use, instead being responsible for an undesired score boost when present. However, this matter is far from resolved, and it would be interesting to see what scoring metrics could improve our ability to study this matter.

3.2 The Effect of Parentheticals

The removal of parentheticals and only parentheticals resulted in the best performing version of the parser. As such, it outperformed the default corpus by about half a percent on shorter sentences, and by an average of a full percent on all sentences of all lengths. Again, there are several alternative conclusions that one can draw from this, and the ambiguity would be resolved by an improved scoring metric, as the one described in 3.1. While it is possible that the inclusion of parentheticals gave the parser some clues as to the location of phrasal boundaries, the opposite could also be the case; the parentheticals could have added noise into an already noisy system. Unlike interjections, parentheticals contain a wide variety of words in a diversity of word orders and lengths.

Table 5. The 40 Most Common Parentheticals

Phrase	# of INTJs	% of INTJs	Phrase	# of INTJs	% of INTJs
you know	431	37.02	I think it was	3	0.257
I mean	105	9.020	I would think	3	0.257
I think	86	7.388	it seems	3	0.257

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

(shaded area composes more than 65% of all cases)

The difference between the inclusion of parentheticals and of other disfluencies, however, is interesting to note. This difference could purely be a result of INTJ/UH information being easy to guess, and PRN information being more difficult, as the former contains words from a small specialized set (uh, um, sure), and the latter is much more akin to intentional speech. This is apparent from the difference between the falloff in the counts in Figures 4 and 5. This difference is relevant in our ability to accurately deduce the label and internal structure of an INTJ and a PRN, and our ability to do so is currently included in our scoring metric. However, our accuracy in these deductions is not as relevant to the extraction of meaning from a sentence as our ability to parse the intentional elements of speech. Modulo this concern, we can conclude that interjections

are marginally useful in parsing, adding a smidgeon (0.1%) to our performance, and that parentheticals are fairly harmful, taking a full percent away.

3.3 The Effect of Sentence Length

In every corpus modification implemented, the parser had a higher success rate on shorter sentences than it did on long ones. This is not at all surprising, as there are fewer possible parses for these shorter sentences. Differences across corpora, however, were more interesting. The extraction of all disfluencies, for example, improved average performance over all sentence lengths relative to the raw corpus, but performance on sentences of length less than forty went down more than five smidgeons. This suggests that the utility of disfluencies in parsing varies from being destructive to constructive depending on sentence length.

3.4 Conclusions

No generalization of disfluency utility holds, as our best performance resulted from the inclusion of some errors and the exclusion of others. We can conclude, however, that the variation between forms of disfluency affects the impact that have on the difficulty of the phrases and sentences in which they exist. Within this variation, we have predictable disfluencies (INTJ/UH), which work to improve the performance of the parser slightly, and relatively unpredictable disfluencies (PRN), whose presence hurts the scores of the sentences in which they exist. We can conclude that a parser of transcribed speech that is able to detect and remove parentheticals before training and evaluation would perform better (by our scoring metric) than a similar parser that cannot. We can also expect that

there is some utility in other forms of disfluency, for the computer listener and perhaps for the human listener.

REFERENCES

1. E. CHARNIAK. *A maximum-entropy-inspired parser.* In *Proceedings of the International Conference on Machine Learning (ICML 2000).* 2000.
2. E. CHARNIAK AND M. JOHNSON. *Edit Detection and Parsing for Transcribed Speech*. In *NAACL 2001*.
3. J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings IEEE Conference on Acoustics, Speech and Signal Processing,* volume I, pages 517-520, San Francisco, March 1992.
4. ROARK, BRIAN. Ph.D. thesis, department of Cognitive Science, Brown University, Providence, RI, 2001.
5. SHRIBERG, E. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, department of Psychology, University of California, Berkeley, CA, 1994. *Conference on Acoustics, Speech and Signal Processing,* volume II, pp. 1005-1008, Atlanta, GA, 1996.
6. STOLCKE, A. AND SHRIBERG, E. *Automatic linguistic segmentation of conversational speech.* In *Proceedings IEEE Conference on Acoustics, Speech and Signal Processing,* volume II, pp. 1005-1008, Atlanta, GA, 1996.

7. STOLCKE, A. AND SHRIBERG, E. *Statistical Language Modeling for Speech Disfluencies.* In *Proceedings IEEE Conference on Acoustics, Speech and Signal Processing,* volume I, pp. 405-409, Atlanta, GA, 1996.