

Parsing Ungrammatical Input: An Evaluation Procedure

Jennifer Foster

Computational Linguistics Research Group
Department of Computer Science
Trinity College, University of Dublin, Ireland
jfoster@tcd.ie

Abstract

This paper presents a procedure for evaluating a parser's ability to produce an accurate parse for an ungrammatical sentence. It is based on the existence of a corpus of ungrammatical sentences, and a parallel corpus containing corrected, and hence grammatical, versions of the sentences in the first corpus. This procedure is applied to a wide-coverage probabilistic parser (Charniak, 2000), and the performance of this parser with respect to ungrammatical input is analysed.

1. Introduction

A usual method for evaluating the accuracy of a natural language parser is to compare the parses produced by the parser for some set of test sentences to gold standard parses for the same set. In this paper an evaluation method is presented which is similar to the above yet which is specifically designed to measure a parser's accuracy in the face of input which is ungrammatical. Section 2. provides an overview of the evaluation method. In Section 3., some of the issues which arise when evaluating a parse for an ungrammatical sentence are discussed. In Section 4., the results of applying the evaluation method to parses produced by a popular probabilistic parser are provided and discussed. Section 5. summarizes the work presented in the previous three sections and presents plans for future work.

2. The Evaluation Procedure: an Overview

The evaluation method is based on three assumptions. The first assumption is that for every ungrammatical sentence there is a grammatical counterpart which expresses the same meaning as the ungrammatical sentence and which would have been produced had the source of error been removed. The second assumption is that a parse for a particular sentence reflects that sentence's meaning. The third assumption, which follows on from the first two—presumably uncontroversial—assumptions, is that the parse for an ungrammatical sentence should be as close as possible to the parse for its grammatical counterpart so that its true meaning is expressed.

A 20,000 word corpus of ungrammatical English sentences from a variety of written language sources (newspapers, emails, academic papers, websites, etc.) has been collected (Foster and Vogel, 2004). Each ungrammatical sentence in the corpus is corrected, producing a parallel corpus of grammatical sentences. For 20% of the ungrammatical sentences, there was more than one correction expressing the same meaning as the ungrammatical sentence. Consider, for example, the ungrammatical clause

*When it is the main method or when you want to take something **from** that could be in the main method and put it into a separate method in the application class*

This could be corrected in two ways, either by deleting the preposition *from* to yield the grammatical

When it is the main method or when you want to take something that could be in the main method and put it into a separate method in the application class

or it could be corrected by deleting the words *that could be in*, yielding the equally grammatical and contextually¹, synonymous

When it is the main method or when you want to take something from the main method and put it into a separate method in the application class

When there are more than one grammatical corrections for a given sentence, all corrected versions are added to the grammatical corpus.

The test data for the evaluation procedure are the ungrammatical sentences in the first corpus. The corpus of corrected grammatical sentences provides the gold standard for the evaluation method. The general evaluation procedure is as follows:

1. For each grammatical sentence
 - (a) Parse sentenceThese are the gold standard parses.
2. For each ungrammatical sentence
 - (a) Parse sentence
 - (b) For each grammatical counterpart
 - Compare parse for ungrammatical sentence with parse for grammatical counterpart. Compute a score for the ungrammatical sentence parse based on its similarity to the grammatical sentence parse.
 - (c) Choose highest score achieved by the ungrammatical sentence parse.

So if an ungrammatical sentence has 2 possible corrections, and its parse is 93% similar to the first correction's parse

¹The ungrammatical sentences are all encountered *in context*, which makes their meaning more transparent. If the meaning of the ungrammatical sentence is not obvious, the sentence is not included in the corpus.

and 89% similar to the second correction's parse, it will receive a score of 93%.

A difficulty associated with gold standard parse sets such as the Penn Treebank (Marcus et al., 1993) is that they are unsuitable for evaluating parses whose structure is not directly comparable with the structure of the treebank parses. This is not an issue for the evaluation method outlined above because the gold standard parses are produced by the same parser as the parses we are evaluating. Thus, this method is general enough to be used with any kind of parse structure.

A generalized version of the PARSEVAL measures of labelled precision and recall is used to compute similarity (Black et al., 1991; Musillo and Sima'an, 2002). These are as follows:

- Precision is $\frac{\#(\text{constituents in grammatical sentence parse} \cap \text{constituents in ungrammatical sentence parse})}{\#(\text{constituents in ungrammatical sentence parse})}$
- Recall is $\frac{\#(\text{constituents in grammatical sentence parse} \cap \text{constituents in ungrammatical sentence parse})}{\#(\text{constituents in grammatical sentence parse})}$

The flexibility in these definitions lies in deciding how to define a constituent and how to calculate the intersection of constituents, and this will, to an extent, depend on the syntactic representations produced by the parser under evaluation.

3. Relevant Issues

Of course there is an obvious difference between the general evaluation of a parser's accuracy and the evaluation of its accuracy with respect to ungrammatical sentences: in the former case, the two parses being compared reflect the same sentence whereas in the latter case the two parses reflect slightly different sentences (since the gold standard parse(s) is the corrected version(s) of the parse under evaluation). The comparison metric needs to take this into account so that the ungrammatical sentence isn't penalized just because it consists of a slightly different set of words to its grammatical counterpart(s). In fact, the comparison metric should allow a parse for an ungrammatical sentence to attain 100% on its precision and recall scores. This will be illustrated in the following sections, with each section describing a particular type of ill-formed sentence.

3.1. Incorrect word form error

Consider, for example, the ungrammatical sentence:

*A romance **in** coming your way.*

which contains a common error involving the mistyping of *is* to produce *in*. The corrected version of this sentence is the grammatical:

A romance is coming your way.

Given the following parse (phrase-structure is depicted using labelled bracketing) for the grammatical sentence:

```
(S (NP a romance)
  (VP is
    (VP coming
      (NP your way))))
```

the following parse for the ungrammatical sentence should be considered completely accurate:

```
(S (NP a romance)
  (VP in
    (VP coming
      (NP your way))))
```

It makes the crucial recognition that the preposition is part of a verb phrase and contrasts in this way with the following less accurate parse for the same sentence:

```
(S (NP (NP a romance)
  (PP in
    (NP (VP coming
      (NP your way))))))
```

3.2. Extraneous word

As a second example consider the case where the ungrammatical sentence contains a superfluous word:

*Annotators parse **to** the sentences in a corpus.*

The corrected version of this sentence is:

Annotators parse the sentences in a corpus.

Given the following parse for the corrected sentence:

```
(S (NP annotators)
  (VP parse
    (NP (NP the sentences)
      (PP in
        (NP a corpus))))))
```

an accurate parse for the ungrammatical sentence would be:

```
(S (NP annotators)
  (VP parse to
    (NP (NP the sentences)
      (PP in
        (NP a corpus))))))
```

where the superfluous *to* does not affect the constituent structure of the sentence. This can be seen more clearly if it is contrasted with another possible parse where *to the sentences in a corpus* is diagnosed as a prepositional phrase:

```
(S (NP annotators)
  (VP parse
    (PP to
      (NP (NP the sentences)
        (PP in
          (NP a corpus))))))
```

3.3. Omitted word

As a third example, consider the erroneous sentence

*Total revenues are expected **to about** EUR 1.6 billion.*

where the infinitival verb *be* has been omitted. A suitable parse of the grammatical version of this sentence would be the following:

```
(S (NP Total revenues)
  (VP are
    (VP expected
      (S (VP to
        (VP be
          (NP about EUR 1.6 billion)
        ))))))))
```

Correspondingly, a completely accurate parse of the ungrammatical sentence would be one with the same phrase structure as the parse for the corrected sentence:

```
(S (NP Total revenues)
  (VP are
    (VP expected
      (S (VP to
          (VP (NP about EUR 1.6 billion)
              )))))
```

In contrast, another plausible parse for the ungrammatical sentence would receive a lower score because it misdiagnoses the infinitival marker *to* as the head of a prepositional phrase:

```
(S (NP Total revenues)
  (VP are
    (VP expected
      (PP to
        (NP about EUR 1.6 billion)
          )))
```

4. Applying the evaluation measure

4.1. Procedure

The evaluation method was applied to a wide-coverage probabilistic parser² (Charniak, 2000). This parser was chosen for two reasons: firstly, because it returns a parse for all the ungrammatical sentences in the corpus which means that its ability to handle ungrammaticality can be meaningfully evaluated using the full corpus, and secondly, because it achieves a high score when evaluated in the standard way using a section of the Penn Treebank. This is important because the evaluation method described here makes the assumption that the parser is able to accurately parse grammatical input. I return to this issue in Section 4.3. below.

The parses produced by Charniak's parses are in the form of phrase-structure trees and so a tree-comparison metric is applied, whereby two phrase-level constituents are the same if they are labelled by the same part-of-speech category and if they enclose the same word sequence. This metric is adapted so that it takes into account the issues described in Section 3.

4.2. Results

The parser achieved a precision score of 91% and a recall score of 91%. In 32% of cases, there was a complete match between the parse for an ungrammatical sentence and the parse for (one of) its grammatical counterparts. The percentage of problematic cases was also calculated: a parse for an ungrammatical sentence was deemed to be problematic if it achieved a precision or recall value of less than 75%. 16% of ungrammatical sentences fell into this category.

Fig. 1 shows the precision/recall scores for each particular type of error occurring in the corpus, along with the complete match and problematic case percentages. For each error type, its frequency in the corpus is indicated as a percentage in brackets beside the name of the error type. A composite error involves two or more of the three main error types, e.g. the ungrammatical clause

It is also worth to remark that....

which is a combination of an incorrect word form (*remark* instead of *remarking*) and an extra word (*to*). This is distinguished from a sentence which contains two or more individual errors, e.g. the sentence

The following roadmap for the has been derived derived for this presentation.

The errors involving an incorrect word form constitute the largest category. The results for the most common type of errors within this category are shown in Fig. 2.

4.3. Discussion

Overall, Charniak's parser has performed well on the ungrammatical sentences in the corpus, achieving 100% accuracy for nearly one third of the test data. For some error types (agreement errors and the use of the wrong preposition), over 70% of cases obtained a complete match, suggesting that these kinds of errors do not tend to affect this parser in a negative way. Here is an example, however, of a sentence from the corpus where the agreement error in the sentence does cause it to be misparsed:

On-going dialogues between the user and the simulated computer system is recorded.

This ungrammatical sentence receives the following parse³:

```
(S (S (NP On-going)
  (VP VBZ dialogues
    (PP between (NP the user))))
  and
  (S (NP the simulated computer system)
    (VP is (VP recorded))))
```

Its corrected version receives the quite different parse:

```
(S (NP (NP (NP On-going dialogues)
  (PP between (NP the user)))
  and
  (NP the simulated computer system))
  (VP are (VP recorded)))
```

The problematic cases increased to over 20% when the ungrammatical sentence contained an erroneous word of a different category to the one it should have been. A typical example is the ill-formed:

Write and tell be all your news.

Charniak's parser produces the following parse for this sentence:

```
(S (VP Write and tell)
  (VP be (NP all your news)))
```

Evaluated against the parse for its grammatical counterpart:

```
(S (VP Write
  and
  tell
  (NP me)
  (NP all your news)))
```

²Downloaded from <ftp://ftp.cs.brown.edu/pub/nlparser/> in March 2003

³Individual word tags have been omitted since they are not taken into account in the evaluation process.

Error Type(% in corpus)	Precision	Recall	Complete Match	Problematic
Incorrect Word Form (44)	91	92	49	15
Missing Word (28)	92	88	19	18
Extra Word (15)	89	93	24	10
Composite Errors (6)	89	89	8	17
More Than One Error (7)	88	89	18	21

Table 1: Results according to error type

Error Type	Precision	Recall	Complete Match	Problematic
Agreement Errors	94	94	74	11
Wrong Preposition	95	95	72	3
Wrong Verb Form	93	94	43	7
Misspelling with Category Change	86	87	24	27

Table 2: Incorrect word forms

3 of its 5 constituents are correct. Similarly, over 20% of sentences containing more than one error are deemed problematic, which is not unexpected.

As mentioned in Section 4.1., the evaluation procedure described in this paper takes a leap of faith by assuming that the parse for the grammatical sentence is correct. Unfortunately, this isn't the case all the time. Charniak's parser, for example, misparses the grammatical clause:

Michelle P probably without significant other

by treating *P* as a verb. The ill-formed clause:

Michelle P probably with out significant other

does not suffer the same problem and *P* is recognized as part of a name. To overcome this problem, each grammatical sentence parse will need to be examined and excluded from the evaluation procedure if it has been misparsed.

5. Conclusion

This paper has presented a procedure for evaluating a parser's ability to produce accurate parses for ungrammatical sentences. The corpus of ungrammatical language collected by the author provides the test sentences and the gold standard parses are provided by the parses produced by the parser under evaluation for the corrected versions of the test sentences. This procedure has been applied to a wide-coverage probabilistic parser (Charniak, 2000). It is hoped to evaluate more parsers using this procedure, e.g. another popular probabilistic Penn-Treebank trained parser (Collins, 1997). It would certainly be interesting to use this procedure to evaluate a parser which returns linguistic structures informationally richer than the trees returned by Charniak's parser, e.g. the typed feature structures returned by the LKB parser (Copestake, 2002).

Acknowledgments

I am grateful to the TCD Broad Curriculum Fellowship for funding this research, and to the anonymous reviewers and Dr. Carl Vogel for their helpful comments.

6. References

- Black, E., S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski, 1991. A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the Speech and Natural Language Workshop*. DARPA.
- Charniak, 2000. A maximum-entropy-inspired parser. In *Proceedings of the NAACL-2000*.
- Collins, Michael, 1997. Three generative lexicalized models for statistical parsing. In *Proceedings of the 35th ACL*. Madrid.
- Copestake, Ann, 2002. *Implementing Typed Feature Structure Grammars*. CSLI Lecture Notes. Cambridge: Cambridge University Press.
- Foster, Jennifer and Carl Vogel, 2004. Good reasons for noting bad grammar: Constructing a corpus of ungrammatical language. In Stephan Kepser and Marga Reis (eds.), *Pre-Proceedings of the International Conference on Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*. Tübingen: organized by the Sonderforschungsbereich 441 "Linguistic Data Structures", University of Tübingen, Germany.
- Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz, 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330.
- Musillo, Gabriele and Khalil Sima'an, 2002. Towards comparing parsers from different linguistic frameworks: An information theoretic approach. In *Proceedings of the "Beyond Parseval - Towards Improved Evaluation Measures for Parsing Systems" Workshop, 3rd LREC*. Las Palmas, Gran Canaria.