The effect of correcting grammatical errors on parse probabilities

Joachim Wagner
CNGL
School of Computing
Dublin City University, Ireland
jwagner@computing.dcu.ie

Jennifer Foster
NCLT
School of Computing
Dublin City University, Ireland.
jfoster@computing.dcu.ie

Abstract

We parse the sentences in three parallel error corpora using a generative, probabilistic parser and compare the parse probabilities of the most likely analyses for each grammatical sentence and its closely related ungrammatical counterpart.

1 Introduction

The syntactic analysis of a sentence provided by a parser is used to guide the interpretation process required, to varying extents, by applications such as question-answering, sentiment analysis and machine translation. In theory, however, parsing also provides a grammaticality judgement as shown in Figure 1. Whether or not a sentence is grammatical is determined by its parsability with a grammar of the language in question.

The use of parsing to determine whether a sentence is grammatical has faded into the background as hand-written grammars aiming to describe only the grammatical sequences in a language have been largely supplanted by treebank-derived grammars. Grammars read from treebanks tend to overgenerate. This overgeneration is unproblematic if a probabilistic model is used to rank analyses and if the parser is not being used to provide a grammaticality judgement. The combination of grammar size, probabilistic parse selection and smoothing techniques results in high robustness to errors and broad language coverage, desirable properties in applications requiring a syntactic analysis of any input, regardless of noise. However, for applications which rely on a parser's ability to distinguish grammatical sequences from ungrammatical ones, e.g. grammar checkers, overgenerating grammars are perhaps less useful as they fail to reject ungrammatical strings.

A naive solution might be to assume that the probability assigned to a parse tree by its probabilistic model could be leveraged in some way to determine the sentence's grammaticality. In this paper, we explore one aspect of this question by using three parallel error corpora to determine the effect of common English grammatical errors on the parse probability of the most likely parse tree returned by a generative probabilistic parser.

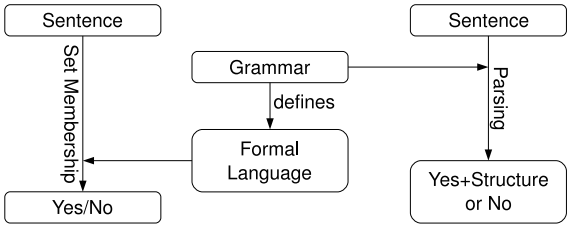

Figure 1: Grammaticality and formal languages

2 Related Work

The probability of a parse tree has been used before in error detection systems. Sun *et al.* (2007) report only a very modest improvement when they include a parse probability feature in their system whose features mostly consist of linear sequential patterns. Lee and Seneff (2006) detect ungrammatical sentences by comparing the parse probability of a possibly ill-formed input sentence to the parse probabilities of candidate corrections which are generated by arbitrarily deleting, inserting and substituting articles, prepositions and auxiliaries and changing the inflection of verbs and nouns. Foster *et al.* (2008) compare the parse probability returned by a parser trained on a regular treebank to the probability returned by the same parser trained on a "noisy" treebank and use the difference to decide whether the sentence is ill-formed.

Research in the field of psycholinguistics has explored the link between frequency and grammaticality, often focusing on borderline acceptable sentences (see Crocker and Keller (2006) for a discussion of the literature). Koonst-Garboden and Jaeger (2003) find a weak correlation between the

frequency ratios of competing surface realisations and human acceptability judgements. Hale (2003) calculates the information-theoretic load of words in sentences assuming that they were generated according to a probabilistic grammar and finds that these values are good predictors for observed reading time and other measures of cognitive load.

3 Experimental Setup

The aim of this experiment is to find out to what extent ungrammatical sentences behave differently from correct sentences as regards their parse probabilities. There are two types of corpora we study: two parallel error corpora that consist of authentic ungrammatical sentences and manual corrections, and a parallel error corpus that consists of authentic grammatical sentences and automatically induced errors. Using parallel corpora allows us to compare pairs of sentences that have the same or very similar lexical content and differ only with respect to their grammaticality. A corpus with automatically induced errors is included because such a corpus is much larger and controlled error insertion allows us to examine directly the effect of a particular error type.

The first parallel error corpus contains 1,132 sentence pairs each comprising an ungrammatical sentence and a correction (Foster, 2005). The sentences are taken from written texts and contain either one or two grammatical errors. The errors include those made by native English speakers. We call this the Foster corpus. The second corpus is a learner corpus. It contains transcribed spoken utterances produced by learners of English of varying L1s and levels of experience in a classroom setting. Wagner et al. (2009) manually corrected 500 sentences of the transcribed utterances, producing a parallel error corpus which we call Gonzaga 500. The third parallel corpus contains 199,600 sentences taken from the British National Corpus and ungrammatical sentences produced by introducing errors of the following five types into the original BNC sentences: errors involving an extra word, errors involving a missing word, real-word spelling errors, agreement errors and errors involving an incorrect verbal inflection.

All sentence pairs in the three parallel corpora are parsed using the June 2006 version of the first-stage parser of Charniak and Johnson (2005), a lexicalised, generative, probabilistic parser achieving competitive performance on Wall Street Journal text. We compare the probability of the highest ranked tree for the grammatical sentence in the pair to the probability of the highest ranked tree for the ungrammatical sentence.

4 Results

Figure 2 shows the results for the Foster corpus. For ranges of 4 points on the logarithmic scale, the bars depict how many sentence pairs have a probability ratio within the respective range. For example, there are 48 pairs (5th bar from left) for which the correction has a parse probability which is between 8 and 12 points lower than the parse probability of its erroneous original, or, in other words, for which the probability ratio is between e^{-12} and e^{-8}. 853 pairs show a higher probability for the correction vs. 279 pairs which do not. Since the probability of a tree is the product of its rule probabilities, sentence length is a factor. If we focus on corrections that do not change the sentence length, the ratio sharpens to 414 vs. 90 pairs. Ungrammatical sentences do often receive lower parse probabilities than their corrections.

Figure 3 shows the results for the Gonzaga 500. Here we see a picture similar to the Foster corpus although the peak for the range from $e^0 = 1$ to $e^4 \approx 54.6$ is more pronounced. This time there are more cases where the parse probability drops despite a sentence being shortened and vice versa. Overall, 348 sentence pairs show an increased parse probability, 152 do not. For sentences that stay the same length the ratio is 154 to 34, or 4.53:1, for this corpus which is almost identical to the Foster corpus (4.60:1).

How do these observations translate to the artificial parallel error corpus created from BNC data? Figure 4 shows the results for the BNC data. In order to keep the orientation of the graph as before, we change the sign by looking at decrements instead of increments. Also, we swap the keys for shortened and lengthened sentences. Clearly, the distribution is wider and moved to the right. The peak is at the bar labelled 10. Accordingly, the ratio of the number of sentence pairs above and below the zero line is much higher than before (overall 32,111 to $167,489 = 5.22$, for same length only 8,537 to 111,171 = 13.02), suggesting that our artificial errors might have a stronger effect on parse probability than authentic errors. Another possible explanation is that the BNC data only contains five error types, whereas the range of

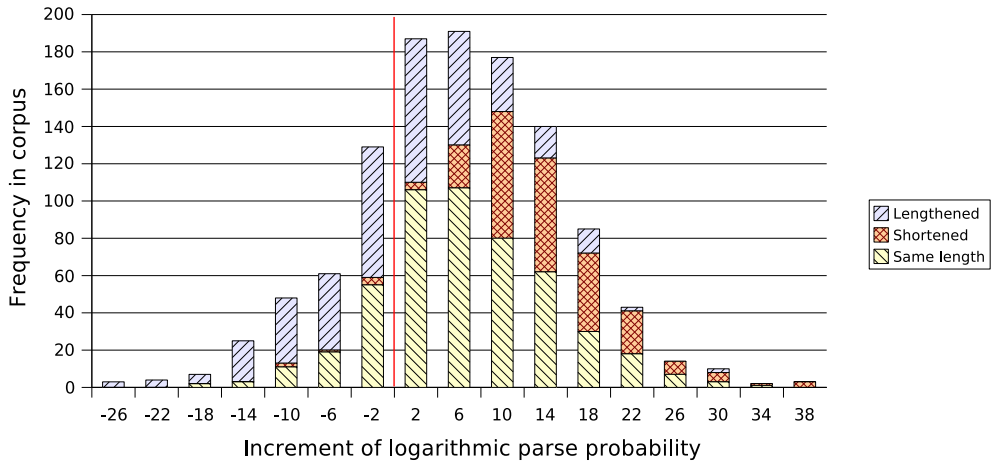

Figure 2: Effect of correcting erroneous sentences (Foster corpus) on the probability of the best parse. Each bar is broken down by whether and how the correction changed the sentence length in tokens. A bar labelled x covers ratios from e^{x-2} to e^{x+2} (exclusive).

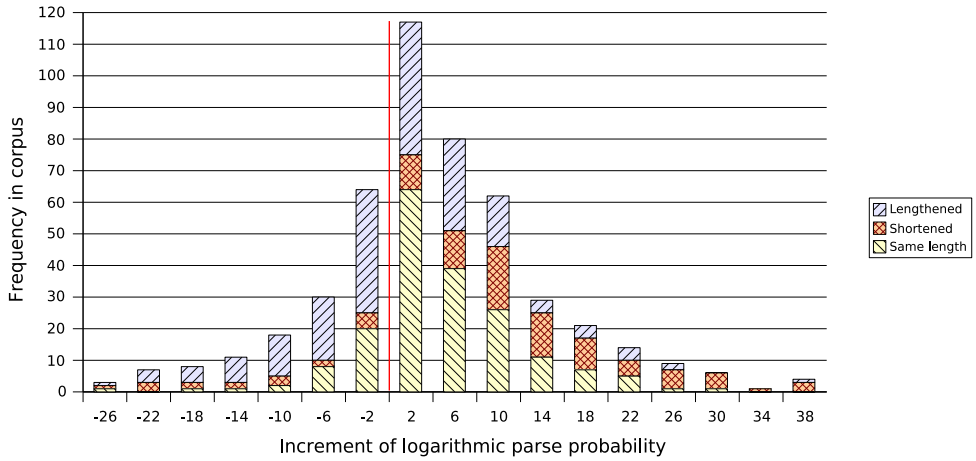

Figure 3: Effect of correcting erroneous sentences (Gonzaga 500 corpus) on the probability of the best parse.

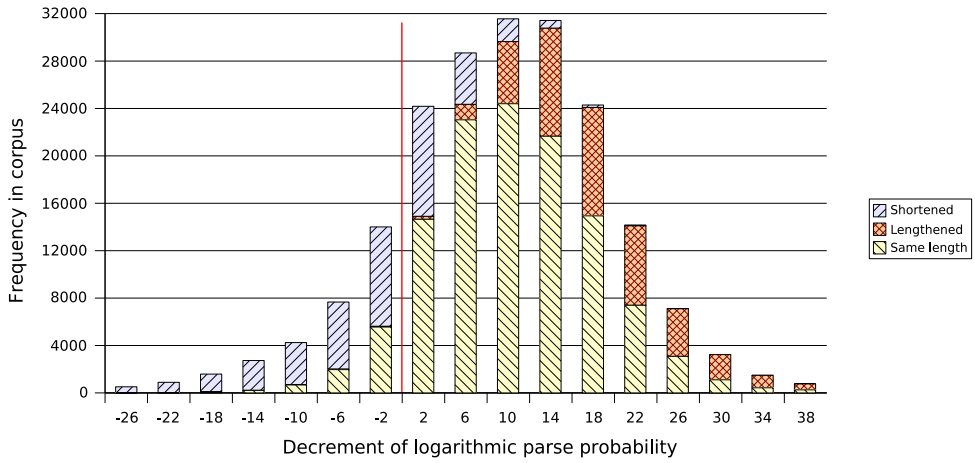

Figure 4: Effect of inserting errors into BNC sentences on the probability of the best parse.

errors in the Foster and Gonzaga corpus is wider.

Analysing the BNC data by error type and looking firstly at those error types that do not involve a change in sentence length, we see that:

- 96% of real-word spelling errors cause a reduction in parse probability.
- 91% of agreement errors cause a reduction in parse probability. Agreement errors involving articles most reliably decrease the probability.
- 92% of verb form errors cause a reduction. Changing the form from present participle to past participle[1] is least likely to cause a reduction, whereas changing it from past participle to third singular is most likely.

The effect of error types which change sentence length is more difficult to interpret. Almost all of the extra word errors cause a reduction in parse probability and it is difficult to know whether this is happening because the sentence length has increased or because an error has been introduced. The errors involving missing words do not systematically result in an increase in parse probability – 41% of them cause a reduction in parse probability, and this is much more likely to occur if the missing word is a function word (article, auxiliary, preposition).

Since the Foster corpus is also error-annotated, we can also examine its results by error type. This analysis broadly agrees with that of the BNC data, although the percentage of ill-formed sentences for which there is a reduction in parse probability is generally lower (see Fig. 2 vs. Fig. 4).

5 Conclusion

We have parsed the sentences in three parallel error corpora using a generative, probabilistic parser and examined the parse probability of the most likely analysis of each sentence. We find that grammatical errors have some negative effect on the probability assigned to the best parse, a finding which corroborates previous evidence linking sentence grammaticality to frequency. In our experiment, we approximate sentence probability by looking only at the most likely analysis – it might be useful to see if the same effect holds if we sum over parse trees. To fully exploit parse or sentence probability in an error detection system, it is necessary to fully account for the effect on probability of 1) non-structural factors such as sentence length and 2) *particular* error types. This study represents a contribution towards the latter.

[1]This raises the issue of covert errors, resulting in grammatical sentence structures. Lee and Seneff (2008) give the example *I am **prepared** for the exam* which was produced by a learner of English instead of *I am **preparing** for the exam*. These occur in authentic error corpora and cannot be completely avoided when automatically introducing errors.

Acknowledgements

We are grateful to James Hunter from Gonzaga University for providing us with a learner corpus. We thank Josef van Genabith and the reviewers for their comments and acknowledge the Irish Centre for High-End Computing for the provision of computational facilities. The BNC is distributed by Oxford University Computing Services.

References

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best-parsing and maxent discriminative reranking. In *Proceedings of ACL*.

Matthew W. Crocker and Frank Keller. 2006. Probabilistic grammars as models of gradience in language processing. In Gisbert Fanselow, C. Féry, R. Vogel, and M. Schlesewsky, editors, *Gradience in Grammar: Generative Perspectives*, pages 227–245. Oxford University Press.

Jennifer Foster, Joachim Wagner, and Josef van Genabith. 2008. Adapting a WSJ-trained parser to grammatically noisy text. In *Proceedings of ACL*.

Jennifer Foster. 2005. *Good Reasons for Noting Bad Grammar: Empirical Investigations into the Parsing of Ungrammatical Written English*. Ph.D. thesis, University of Dublin, Trinity College.

John Hale. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123.

Andrew Koontz-Garboden and T. Florian Jaeger. 2003. An empirical investigation of the frequency-grammaticality correlation hypothesis. Student essay received or downloaded on 2006-03-13.

John Lee and Stephanie Seneff. 2006. Automatic grammar correction for second-language learners. In *Interspeech 2006 - 9th ICSLP*, pages 1978–1981.

John Lee and Stephanie Seneff. 2008. Correcting misuse of verb forms. In *Proceedings of ACL*.

Guihua Sun, Xiaohua Liu, Gao Cong, Ming Zhou, Zhongyang Xiong, John Lee, and Chin-Yew Lin. 2007. Detecting erroneous sentences using automatically mined sequential patterns. In *Proc. of ACL*.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2009. Judging grammaticality: Experiments in sentence classification. *CALICO Journal*, 26(3).