The Internet as a Normative Corpus: Grammar Checking with a Search Engine

Jonas Sjöbergh
KTH Nada
SE-100 44 Stockholm, Sweden
jsh@nada.kth.se

Abstract

In this paper some methods using the Internet as a normative corpus for error checking purposes is presented. These include error detection and removing false alarms from existing grammar checkers. We evaluate these methods on Swedish texts. While not performing as well as state of the art traditional methods, results indicate that these methods are still useful, especially as a complement to other methods. Errors not detected by traditional methods can be detected by very simple means, and increasing the precision of other grammar checkers by removing false alarms also works quite well.

1 Introduction

Currently there is a lot of research in natural language processing based on the idea that you can get good results using simple methods as long as you have very large amounts of data, often outperforming more sophisticated methods using smaller amounts of data.

The Internet is a large and freely available corpus, so it is appealing to use it for different purposes. Some work using similar approaches to our own include estimating bigram frequencies for rare bigrams (Keller and Lapata, 2003), suggesting improvements on text constructions where the author is unsure (Moré et al., 2004) and detecting malapropisms (Bolshakov, 2005).

When using the Internet as an example of correct language use, as we do here, there are some problems. There are many web sites with intentional examples of incorrect language use, and recognizing these can be hard. Publishing text on the Internet is cheap and easy, with no requirements regarding proofreading, so there are also many unintentional errors. These problems are not that bad in practice, since there are usually more examples of correct constructions than the corresponding erroneous constructions. As long as the possibility of errors is taken into account, many methods using the Internet as a normative corpus work quite well.

Word	Internet pages	Parole occurrences
välde	4 190	33
multnade	121	1
ett	3 710 000	139 766
den	5 080 000	199 223

Table 1: Occurrences in a 20 million words corpus and using an Internet search engine.

Another problem is that while the Internet is large, it is too small for many interesting ideas. This is harder to deal with, but the Internet is still growing quite fast, so just by waiting more and more data is made available.

2 Internet size

When using the Internet as a large corpus it is interesting to know roughly how large it is. Since it grows all the time there is no official size available. The size also varies depending on which search engine (or other method) you use to access it.

We used the search engine `eniro.se` in this paper. While other search engines give access to more documents, this one has some advantages. The output is very easy to parse, there is no limit on the number of searches each day and it has an "only pages in Swedish" option, which was useful since we evaluated the methods on Swedish texts.

Method	Genre	Limit	Correct	False
Internet	newspaper	10 000	1	21
Internet	newspaper	100	2	162
Granska	newspaper	-	8	35
MS Word	newspaper	-	10	92
Internet	learner	10 000	21	4
Internet	learner	100	100	22
Internet	learner	0	283	27
Granska	learner	-	411	13
MS Word	learner	-	392	21

Table 2: Using word bigrams to detect errors, in newspaper texts and second language learner essays. "Limit" is the minimum number of occurrences on the Internet of each word required to try the bigram lookup. Granska and MS Word are two state of the art grammar checkers included for comparison.

Using the Internet search engine `eniro.se` with the "only pages in Swedish" option enabled we searched for a few words chosen more or less at random. Some relatively rare words, which probably occur only a few times on each web page, and some common words, which probably occur many times on each page. We then compared the number of pages returned by the search engine to the number of occurrences of the words in the Swedish Parole corpus (Gellerstam et al., 2000).

For the rare words there were about 100 times more pages than occurrences in the corpus. For common words there were about 25 times more pages than occurrences in the corpus, see Table 1. This difference between common and rare words is of course caused by the common words occurring many times on each page in the search engine index.

The Swedish Parole contains 20 million words, so a low estimate would give a few billion words of Swedish indexed by this search engine. Swedish is a relatively large language on the Internet (though not very large in the number of speakers). English is of course the number one language on the Internet, with a very large margin to language number two.

These numbers give a rough idea of what sort of statistics are reasonable to collect. For instance word trigram occurrences would not be reasonable, since even a low estimate of 100 000 possible word forms would lead to very sparse data indeed, even for English. In the next section we would like to use occurrences of n-grams of words, but even for bigrams the data will be sparse, especially for rare words.

3 Detecting Errors

ProbGranska (Bigert and Knutsson, 2002) is an existing grammar checker that detects unlikely part-of-speech trigrams, trained on a corpus of correct text. Inspired by this we used a similar idea, using words instead of PoS tags. All word bigrams in a text were sent to a search engine. Bigrams not occurring on the Internet were reported as errors. We tried this on newspaper texts and on essays written by learners of Swedish. This found spelling errors, erroneously split compounds, agreement errors, missing words and more. Results can be seen in Table 2.

We compared the results to two state of the art grammar checkers, the one included in the Swedish version of MS Word 2000 (Arppe, 2000; Birn, 2000) and Granska (Domeij et al., 2000). Both are based on manually written rules for different error types. They of course outperform our method, mostly because they detect a lot of spelling errors but also because they detect errors using a larger scope than our method.

Since the Internet is too small for good coverage of Swedish word bigrams there are many false alarms from our method, especially on the newspaper texts. Checking only bigrams where both words are common mitigates this, but lowers recall. All spelling errors go undetected, for instance.

The performance on newspaper texts is quite bad, but on the other hand there are

	Detections	False Alarms	Precision
Original	102	19	84%
Filtered	84	7	92%

Table 3: Filtering suspected errors from the grammar checker ProbGranska using the Internet. Evaluated on essays written by second language learners.

almost no errors in the text so very few detections can be expected. On the second language learner essays quite good results are achieved, though still worse than state of the art grammar checkers. Learners use a limited vocabulary, mainly common words, which is well covered on the Internet, resulting in few false alarms. Learners also make many errors detectable by this method.

This method only finds very local errors. It also has problems with phrase and clause boundaries and some multi-word expressions, and of course rare words. Some improvements include ignoring numbers, interjections and proper names, which can be identified relatively well with automatic methods.

Data is still very sparse for normal language users, since there are several hundred thousand word forms that are commonly used, and we only have a few billion words of text in our "corpus". This means that a bigram in general has very low probability of occurring on the Internet. This sparseness is still a problem for languages with more text available on the Internet, so even if we would be interested only in English, the problem would remain (though somewhat mitigated).

Other than being very resource lean, our method also has another advantage. Of the 21 detected errors in learner essays checking only common words, 8 errors were not detected by any of four other available grammar checkers, including the two state of the art methods above. When checking only such bigrams the number of false alarms is very low.

This indicates that this method can be used together with other methods. This would improve error coverage while introducing very few new false alarms. So while this simple method for error detections does not work very well in general, it does complement other methods and can work well for certain types of users.

4 Removing False Positives

Another use of the Internet is to remove false positives (false alarms). Instead of letting the lack of certain constructions be an indication that they are wrong, we can use the occurrences of certain constructions to indicate that they are correct.

This can be done by taking the suspected errors from a grammar checker and sending these constructions to a search engine. If these have been used a sufficient number of times on the Internet, treat the suspected error as a false alarm. It is a good idea to require more than one occurrence on the Internet, since there are bound to be some errors, intentional or otherwise, on the Internet. We have tried this for two different grammar checkers. Both are based on automatic methods and thus has a tendency to produce quite a few false alarms, especially on text domains that differ from the training texts.

4.1 ProbGranska

ProbGranska (Bigert and Knutsson, 2002) detects unlikely part-of-speech (PoS) trigrams. This leads to quite a lot of false alarms in general, because the PoS trigram data is quite sparse. ProbGranska already has strategies to mitigate this, it detects phrase and clause boundaries and has some substitution procedures for rare PoS tags.

To increase precision further we used the Internet. ProbGranska points out PoS trigrams as suspected errors. For each such trigram we took the corresponding trigram of words and checked how many web pages contain that trigram. If there were more than 25 hits with the search engine the error was removed as a false alarm.

This gives the filter a shorter scope than the original error detection. The filter only looks at three words, while the tagging step that produces the PoS trigram can look at the neighboring words and their PoS as well.

	Split compounds	Other errors	False alarms
Original	19	27	10
Filtered	16 (0)	3 (19)	0 (3)

Table 4: Filtering suspected split compounds from the grammar checker SnålGranska, in second language learner essays. Numbers in parenthesis are detections which remain but had the diagnosis changed to “other error type”.

When tried on 10 000 words of learner essays precision was increased from 84% to 92%, but quite a few of the correct detections were also removed, see Table 3. On 10 000 words of newspaper texts, 16 of 36 false alarms were removed. Since there were very few errors in these texts, there was only one correct detection. The correct detection was not removed.

4.2 Split Compounds

Split compounds is a quite common error type in Swedish (and other compounding languages, such as German). It is quite hard to detect these errors with automatic methods, and few grammar checkers for Swedish try to handle this error type. There are many (erroneous) split compounds on the web, which means that checking if the suspected error occurs on the Internet is not a very good way to filter false alarms for this error type. Too many correct detections are removed.

For split compounds of Swedish, one can instead combine the words of the suspected split compound into a compound word. If this word exists on the Internet it was a correct alarm, otherwise it was a false alarm.

This removes many false alarms. This also removes detections of errors which are not split compounds but still erroneous. Some error types sometimes look like split compounds, examples include agreement errors and using the wrong word class, such as adjective form instead of adverb, noun instead of verb. It would probably be good for the writer to get an error report on such errors, even if the diagnosis was “split compound”.

Still, it would be better if they were detected with the correct diagnosis, perhaps by a different grammar checker module. If we want a good split compound detection module these should be removed.

It is possible to modify the simple filtering method above to handle such errors better. The words are combined into a compound as before, and as before, if this compound is more common than the multi-word expression we treat it as a correctly detected split compound. If neither the compound nor the multiword expression occurs on the Internet more than 10 times it is probably not a split compound, but it is probably still an error. These detections are given another diagnosis, such as “error, but not a split compound”.

The grammar checker SnålGranska (Sjöbergh and Knutsson, 2005) detects split compound errors (and some other error types). It has quite good recall for these errors compared to other grammar checkers. It has a relatively low precision though, so there is potential for improvement by removing false alarms.

When using the first mentioned method to remove false alarms for split compounds 16 of 29 split compound false alarms are removed on 10 000 words of newspaper text. Using the filter that relabels errors only removes 5 false alarms, while the other 11 are relabeled. There were no correct detections of split compounds in these texts, since there were no errors to detect.

On second language learner essays there are more errors to detect. The filter removes most false alarms and also correctly relabels most errors of other types, see Table 4. 16 of 19 correctly detected split compounds remain, with the correct diagnosis. 3 errors of other error types are still labeled as split compounds and 3 false alarms remain, though no longer believed to be split compounds.

5 Conclusions and discussion

The Internet is a large corpus. This means that it is possible to get interesting results using very simple methods, such as bigram lookup. It is too small for many interesting uses, though.

While the Internet is often too small for normal users it might be large enough for special applications. One example is learners of a new language, who use a limited vocabulary. This vocabulary tends to be common words, which are well covered on the Internet.

The Internet can also be used as a complement to traditional methods, by for instance removing false alarms or detecting some error types missed by other methods.

References

Antti Arppe. 2000. Developing a grammar checker for Swedish. In T. Nordgård, editor, *Proceedings of Nodalida '99*, pages 13–27. Trondheim, Norway.

Johnny Bigert and Ola Knutsson. 2002. Robust error detection: A hybrid approach combining unsupervised error detection and linguistic knowledge. In *Proceedings of Romand 2002, Robust Methods in Analysis of Natural Language Data*, pages 10–19, Frascati, Italy.

Juhani Birn. 2000. Detecting grammar errors with lingsoft's Swedish grammar checker. In T. Nordgård, editor, *Proceedings of Nodalida '99*, pages 28–40. Trondheim, Norway.

Igor Bolshakov. 2005. An experiment in detection and correction of malapropisms through the web. In *Proceedings of CICling 2005*, pages 803–815, Mexico City, Mexico.

Richard Domeij, Ola Knutsson, Johan Carlberger, and Viggo Kann. 2000. Granska – an efficient hybrid system for Swedish grammar checking. In *Proceedings of Nodalida '99*, pages 49–56, Trondheim, Norway.

Martin Gellerstam, Yvonne Cederholm, and Torgny Rasmark. 2000. The bank of Swedish. In *Proceedings of LREC 2000*, pages 329–333, Athens, Greece.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Joaquim Moré, Salvador Climent, and Antoni Oliver. 2004. A grammar and style checker based on internet searches. In *Proceedings of LREC-2004*, pages 1931–1934, Lisbon, Portugal.

Jonas Sjöbergh and Ola Knutsson. 2005. Faking errors to avoid making errors: Very weakly supervised learning for error detection in writing. In *Proceedings of RANLP 2005*, pages 506–512, Borovets, Bulgaria.