Mining the Web for Discourse Markers

Ben Hutchinson

School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, Scotland, UK
B.Hutchinson@sms.ed.ac.uk

Abstract

This paper proposes a methodology for obtaining sentences containing discourse markers from the World Wide Web. The proposed methodology is particularly suitable for collecting large numbers of discourse marker tokens. It relies on the automatic identification of discourse markers, and we show that this can be done with an accuracy within 9% of that of human performance. We also show that the distribution of discourse markers on the web correlates highly with those in a conventional balanced corpus.

1. Introduction

Discourse markers are words or phrases which link clauses or sentences by signalling relations between them, e.g. *because*, *afterwards*, or *assuming that*. This paper is concerned with the automatic collection of sentences containing discourse markers from the World Wide Web. Having large numbers of such sentences is important for both the empirical linguistic study of discourse markers, and also the computer processing of natural language discourse.

The study of discourse markers constitutes an important step in the study of natural language discourse. Historically, the empirical study of discourse markers has involved manual analysis of example sentences (Knott, 1996, for example). The speed and difficulty of manual analysis has meant that claims have been made on the basis of relatively few examples. Recently, however, researchers have begun applying automatic methods of analysis to large corpora. Bestgen et al. (2003) have used Latent Semantic Analysis to investigate the distributions of causal discourse markers. Hutchinson (2003) and Bestgen et al. (2003) have shown that lexical co-occurrences provide evidence of differences and similarities in meaning of discourse markers.

Natural Language Processing has also begun to use large numbers of discourse markers in approaches to discourse understanding and generation. Within discourse understanding, one strain of research has focussed on the task of distinguishing discourse and sentential uses of discourse markers (Siegel and McKeown, 1994; Litman, 1996; Marcu, 1998). Discourse markers have also been used in the unsupervised learning of discourse relations (Marcu and Echihabi, 2002; Lapata and Lascarides, 2004). Within discourse generation, corpora of discourse markers have been used to train statistical models predicting *if* discourse markers should be generated, and if so *where* they should be placed (Di Eugenio et al., 1997).

Large scale empirical research such as this requires many examples of discourse markers to train on. But even a large corpus such as the British National Corpus (BNC) (see Burnard (1995)), with 100 million words, contains few instances of some discourse markers; for example, it contains just nine matches of the string *always assuming that*, and at least two of these matches are not discourse markers. In fact, it has been pointed out that the BNC does not contain enough data for statistically stable conclusions about most English lexical items (Kilgarriff and Grefenstette, 2003).

Marcu and Echihabi's (2002) solution to the sparseness problem is to manually construct a 1 billion word training corpus, by combining previously existing resources. In contrast, this paper proposes overcoming data sparseness by using the web as a source of example sentences. We automatically construct a database of sentences each of which contains a discourse marker.

We proceed by first considering the use of the web as a source of linguistic data. In Section 3 we present our methodology for obtaining discourse markers from the web. In Section 4 the proposal is evaluated. We conclude and discuss future work in Section 5.

2. The web as a source of data

It may be argued that the World Wide Web is unrepresentative of language use in general. However Kilgarriff and Grefenstette (2003) point out that our understanding of what it means to be representative is quite primitive. They conclude: "The web is not representative of anything else. But neither are other corpora, in any well-understood sense." Despite questions about its representativeness, the web is increasingly viewed as a source of valuable linguistic data (for an overview see Kilgarriff and Grefenstette (2003)). Indeed, a web search engine was launched recently aimed specifically at aiding linguistic research (Resnik and Elkiss, 2004).

Furthermore, recent empirical work has shown that the web can be used as a reliable source of certain types of statistical linguistic information. For example, Keller and Lapata (2003) show that the web provides statistics on Adjective-Noun, Noun-Noun and Verb-Object bigrams that correlate well with statistics from both a large balanced corpus (the BNC), and an even larger single domain corpus (the North American News Text Corpus). In addition, they also show that human plausibility judgements correlate better with the bigram statistics from the web than they do with those from either of the conventional corpora. It is therefore an interesting question as to whether similar results can be obtained for discourse markers. We will show below that discourse marker co-occurrences obtained from the web do correlate highly with those from the BNC.

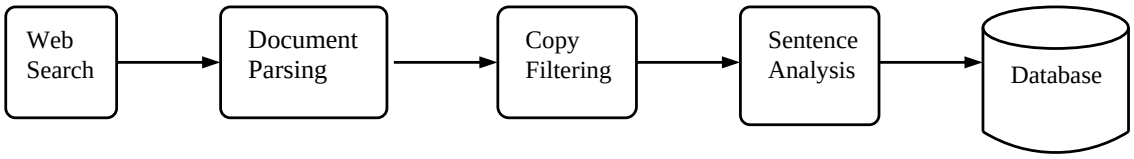

Figure 1: Methodology for mining the web

3. A methodology for mining the web for discourse markers

We shall now describe a methodology for mining the web for sentences containing discourse markers. This task is made difficult by the fact that many discourse markers also have non-discourse uses. For example, *assuming that* is not a discourse marker in *I was assuming that you'd left*, while *and* may coordinate any constituents of the same type.

3.1. System description

Our methodology is shown schematically in Figure 1, and each of the main stages is summarised below.

Step 1: Searching the web

A search engine is used to find pages that may contain discourse markers, by searching for their surface forms. For example, to collect examples of the discourse marker *and*, we begin by doing a web search for `"and"`.

One difficulty is that many search engines restrict how many hits may be accessed per search. For example AltaVista only returns the top 1,000 hits. Our approach to overcoming this is to use digits as additional search terms. For example, using AltaVista we can retrieve 1,000 pages containing both *and* and the digit 1 by searching for `"and" AND 1`. Similarly, we can retrieve 1,000 pages containing *and* but not 1 by searching for `"and" AND NOT 1`. Thus we retrieve a total of 2,000 pages containing *and*. In doing this we make an implicit assumption that the distributions of `"and"` and `1` are independent, but this is unlikely to be harmful.

Step 2: Document parsing

The URLs returned by the search engine are downloaded and analysed automatically. An HTML parser is used to extract textual elements from the document, and punctuation heuristics are used to segment the text into sentences. Sentences not containing strings matching the relevant surface forms are filtered out.

At this stage we will have a list of sentences containing both discourse and non-discourse uses of *and*, for example.

Step 3: Copy filtering

Multiple copies of identical sentences found on the web are discarded. The motivation for this is twofold. Firstly, we do not want to waste processing time by analysing the same sentence repeatedly.

Secondly, we aim to avoid repetitions of a single utterance affecting our statistics. Such repetitions may occur through the mirroring of websites, syndication of news items or columns, plagiary, or quotation. For example, the discourse marker *and* occurs in the sentence

> All programmers are playwrights and all computers are lousy actors.

This sentence scores 1,150 hits on Google, and all these hits probably stem from a single creative utterance. It is repetitions such as these that we want to avoid.

A negative consequence of this decision is that we do not capture the frequency with which the same sentence may be created independently. For example, we ignore the information that *Come and get it!* is a common use of the discourse marker *and*.

Step 4: Sentence analysis

A parser is run on each sentence, and the resulting parse tree is automatically analysed to determine if the previously identified surface forms are actually discourse markers. This approach is novel, in contrast to previous surface-based approaches (Marcu, 1998, for example). Sentences not containing discourse markers are discarded at this stage.

Because the web has the opportunity to provide a huge amount of training examples, we can afford to be conservative in our identification of discourse markers. However we must also be careful that in being conservative we do not collect an unrepresentative sample of data.

Step 5: Database update

Sentences identified as containing discourse markers are saved to a database, indexed by the discourse markers they contain, for later analysis. This indexing makes it easy to use the database as a resource for analysing the distributions of particular discourse markers.

3.2. Implementation

The above procedure has been implemented using the AltaVista search engine, CPAN's HTML::Parser module, and Charniak's (2000) statistical parser. The system was used to mine the web for sentences containing each of a set of 116 discourse markers, by analysing 8,000 pages containing each discourse marker's surface form. These discourse markers were all structural connectives, in that syntactically they function as to combine clauses within a sentence (Webber et al., 2003). For example, the structural connectives included coordinating conjunctions such as *and*, subordinating conjunctions such as *after*, and a variety of multi word expressions with diverse internal compositions, including *assuming that, as long as, every time, given that, despite the fact that, except after, even since, on condition that,* and *to the extent that*.

Discourse marker	In BNC	In 8,000 web pages	Extrapolation to entire web	(#page hits using AlltheWeb)
and	371,43	33,626	5,200,550,000	(1,237,269,101)
after	30,551	9,132	307,068,000	(269,003,783)
as long as	2,357	6,244	19,873,100	(25,461,952)
assuming that	395	4,695	1,460,880	(2,489,247)
every time	660	5,124	9,400,120	(14,676,227)

Table 1: Number of sentences identified as containing discourse markers

In identifying structural connectives automatically, the main requirement we have is that the candidate connective immediately precedes an S node in the parse tree. Figure 2 gives example fragments of parse trees that would be identified as containing discourse markers. Note that the algorithm is robust in the face of some parser errors, such as the unusual expansion `PP→IN S` in the third example. Because the identification of discourse markers is the most difficult step in the process, we include an evaluation of the performance of our module in Section 4.2..

```
(S ...) (CC and) (S...)
(SBAR (IN after) (S...))
(PP (IN after) (S...))
(PP (VBN given) (SBAR (IN that) (S...)))
(NP (DT the) (NN moment) (SBAR...))
(ADVP (RB as) (RB long) (SBAR (IN as) (S...)))
(PP (IN in) (SBAR (IN that) (S...)))
```

Figure 2: Identifying structural connectives

4. Evaluation

Evaluation consisted of three parts: 1) comparing the number of example sentences obtainable from the web with the number available in the BNC, 2) checking the accuracy of the Sentence Analysis module in identifying true discourse markers, and 3) estimating how well sentences obtained from the web are representative of general usage of discourse markers. We now describe each of these in detail.

4.1. Quantity of sentences obtainable from web

The first two stages of the evaluation used five discourse markers that had different syntactic compositions, signalled different discourse relations, and had different frequencies in the BNC. They were *and*, *after*, *assuming that*, *as long as* and *every time*. For each, we compared the numbers of sentences in our database with the number of those obtained from the BNC using the same Sentence Analysis module (see Table 1). Furthermore, we were also able to estimate the total number of tokens of each discourse marker available on the web in February 2004. This was done by using the AlltheWeb search engine to see how many pages contained the surface form, and extrapolating from the number of discourse marker tokens in the 8,000 analysed documents. These figures should be considered lower bounds due to the emphasis on precision in the Sentence Analysis module. The results illustrate the enormous potential of the web as a source of data about the distribution of discourse markers.

4.2. Accuracy of the Sentence Analysis

As mentioned above, in the Sentence Analysis stage we regard precision as more important than recall, and so it is precision that we evaluate here. For each of the same five discourse markers, 100 examples from the database were selected at random. These were inspected by two human judges, who were asked whether each sentence contained the given discourse marker. The results are shown in Table 2. Since the judges disagreed on 7% of cases, an upper bound of 93% can be inferred for the task. The results show that the sentence analysis module achieved an accuracy of between 84.4% and 91.4%, depending on how the judges' disagreements might be resolved. This is comparable to Marcu's (1998) achievement of 89.5% precision on identifying a set of 275 discourse marker tokens.

	Marked correct by:		Inter-judge
	2 judges	1 judge	agreement (K)
and	88.0%	2.0%	0.898
after	80.0%	12.0%	0.505
as long as	88.0%	9.0%	0.363
assuming that	76.0%	8.0%	0.750
every time	90.0%	4.0%	0.729
Total	84.4%	7.0%	0.671

Table 2: Accuracy of Sentence Analysis

There was variation across discourse markers, both in accuracy of the system and in disagreement between human judges. To evaluate the significance of the variation in the system's performance, we converted the results in Table 2 to scores out of 200. For example, on *every time* the system agreed with both judges 90 times and just one judge 4 times, for a score of 184/200. We then applied the χ^2 test. The system's accuracy on *assuming that* differed significantly from its performance on each of *as long as* ($\chi^2 = 13.18; p < 0.001$), *every time* ($\chi^2 = 11.96; p < 0.001$) and *and* ($\chi^2 = 6.18; p < 0.025$). Its performance on *after* also differed from that on *as long as* ($\chi^2 = 4.40; p < 0.05$).

Overall, the two judges agreed 93% of the time. The significance of this was evaluated using the kappa statistic (Carletta, 1996). The kappa statistic is defined as

$$K = \frac{P_A - P_0}{1 - P_0}$$

where P_A is the probability that the judges agree in practice, and P_0 is the probability that they would have agreed by chance. In our case, $K = 0.671$, indicating a good level of agreement, however agreement was poor for *as long as*.

4.3. Quality of sentences mined from web

The web may be claimed to be unrepresentative. Here we attempted to quantify representativeness by comparing discourse marker co-occurrences obtained from the web with those in the BNC. This stage of the evaluation used a different collection of discourse markers than the previous stages. This time only high frequency discourse markers were chosen, as these provide more reliable data for calculating correlation statistics. They are listed in Table 3.

Comparison between the web sentences and the BNC ones was done by comparing bigram counts for co-occurrences of structural connectives and adverbial discourse markers (e.g. *furthermore, then, as a result, afterwards*), along the lines of Hutchinson (2003). These bigrams are indicative of the discourse contexts in which the discourse markers appear. The reason there are more BNC bigrams than Web bigrams, is that these five discourse markers occur with a higher frequency in the BNC than they do in 8,000 web pages containing their surface strings (which may or may not be discourse markers). This is not the case in general. Correlation was measured using Pearson's r, and the results indicate a high degree of correlation. This suggests sentences from the web contain discourse markers which are representative of their discourse contexts.

	Correlation	#BNC bigrams	#Web bigrams
after	0.8028**	1,504	258
and	0.9259**	126,714	11,153
before	0.8555**	4,329	398
but	0.9578**	87,193	7,159
or	0.8898**	2,677	454

$^{**}p < 0.00005$

Table 3: Correlation between discourse marker bigrams from the BNC and the web

These correlation statistics are higher on average than those found by Keller and Lapata (2003) for Adjective-Noun ($r = 0.847$), Noun-Noun ($r = 0.720$) and Verb-Object ($r = 0.762$) bigrams. This may be due to their being fewer distinct discourse markers than there are verbs, nouns or adjectives.

5. Conclusions and future work

We have proposed a procedure for automatically extracting sentences containing discourse markers from the web. We showed that many more examples can be obtained from the web using this procedure than can be obtained in the BNC. Accuracy in identifying discourse markers was high overall, although there was significant variation across discourse markers. Note that in empirical NLP some degree of noise in is acceptable when large amounts of data are available. Finally, the high correlation of discourse marker bigrams shows that example sentences mined from the web are in at least some respects similar to those gathered from the BNC.

In future work we will expand the database by also applying the methodology to gathering adverbial discourse markers.

Acknowledgements

This research was supported by EPSRC Grant GR/R40036/01 and a University of Sydney Travelling Scholarship.

6. References

Bestgen, Yves, Liesbeth Degand, and Wilbert Spooren, 2003. On the use of automatic techniques to determine the semantics of connectives in large newspaper corpora: An exploratory study. In *Proceedings of the MAD'03 workshop on Multidisciplinary Approaches to Discourse*.

Burnard, Lou (ed.), 1995. *Users Reference Guide for the British National Corpus*. Oxford University Computing Service.

Carletta, Jean, 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Di Eugenio, Barbara, Johanna D. Moore, and Massimo Paolucci, 1997. Learning features that predict cue usage. In *Proceedings of the 35th Conference of the Association for Computational Linguistics (ACL97)*. Madrid, Spain.

Hutchinson, Ben, 2003. Automatic classification of discourse markers by their co-occurrences. In *Proceedings of the ESSLLI 2003 workshop on Discourse Particles: Meaning and Implementation*.

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

Kilgarriff, Adam and Gregory Grefenstette, 2003. Introduction to the special issue of the web as corpus. *Computational Linguistics*, 29(3):333–348.

Knott, Alistair, 1996. *A data-driven methodology for motivating a set of coherence relations*. Ph.D. thesis, University of Edinburgh.

Lapata, Mirella and Alex Lascarides, 2004. Inferring sentence-internal temporal relations. In *In Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting*. Boston, MA.

Litman, Diane J., 1996. Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, 5:53–94.

Marcu, Daniel, 1998. A surface based approach to identifying discourse markers and elementary textual units in unrestricted texts. In *Proceedings of the COLING/ACL workshop on Discourse Relations and Discourse Markers*.

Marcu, Daniel and Abdessamad Echihabi, 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*. Philadelphia, PA.

Resnik, Philip and Aaron Elkiss, 2004. The linguist's search engine: Getting started guide. Technical Report LAMP-TR-108/CS-TR-4541/UMIACS-TR-2003-109, University of Maryland, College Park. Update of January 20, 2004.

Siegel, Eric V. and Kathleen McKeown, 1994. Emergent linguistic rules from inducing decision trees: Disambiguating discourse clue words. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*.

Webber, Bonnie, Matthew Stone, Aravind Joshi, and Alistair Knott, 2003. Anaphora and discourse structure. *Computational Linguistics*.