Less is more:
Eliminating index terms from subordinate clauses

Simon H. Corston-Oliver and William B. Dolan
Microsoft Research
One Microsoft Way
Redmond WA 98052
{simonco, billdol}@microsoft.com

Abstract

We perform a linguistic analysis of documents during indexing for information retrieval. By eliminating index terms that occur only in subordinate clauses, index size is reduced by approximately 30% without adversely affecting precision or recall. These results hold for two corpora: a sample of the world wide web and an electronic encyclopedia.

1 Introduction

Efforts to exploit natural language processing (NLP) to aid information retrieval (IR) have generally involved augmenting a standard index of lexical terms with more complex terms that reflect aspects of the linguistic structure of the indexed text (Fagan 1988, Katz 1997, Arampatzis et al. 1998, Strzalkowski et al. 1998, inter alia). This paper shows that NLP can benefit information retrieval in a very different way: rather than increasing the size and complexity of an IR index, linguistic information can make it possible to store *less* information in the index. In particular, we demonstrate that robust NLP technology makes it possible to omit substantial portions of a text from the index without dramatically affecting precision or recall.

This research is motivated by insights from Rhetorical Structure Theory (RST) (Mann & Thompson 1986, 1988). An RST analysis is a dependency analysis of the structure of a text, whose leaf nodes are the propositions encoded in clauses. In this structural analysis, some propositions in the text, called "nuclei," are more centrally important in realizing the writer's communicative goals, while other propositions, called "satellites," are less central in realizing those goals, and instead provide additional information about the nuclei in a manner consistent with the discourse relation between the nucleus and the satellite. This asymmetry has an analogue in sentence structure: main clauses tend to represent nuclei, while subordinate clauses tend to represent satellites (Matthiessen and Thompson 1988, Corston-Oliver 1998).

From the perspective of discourse analysis, the task of information retrieval can be viewed as attempting to identify the "aboutness," or global topicality, of a document in order to determine the relevance of the document as a response to a user's query. Given an RST analysis of a document, we would expect that for the purposes of predicting document relevance, information that occurs in nucleic propositions ought to be more useful than information that occurs in satellite propositions. To test this expectation, we experimented with eliminating from an IR index those terms that occurred in certain kinds of subordinate clauses.

2 System description

At the core of the Microsoft English Grammar (MEG), is a broad-coverage parser that produces conventional phrase structure analyses augmented with grammatical relations; this parser is the basis for the grammar checker in Microsoft Word (Heidorn 1999). Syntactic analyses undergo further processing in order to derive logical forms (LFs), which are graph structures that describe labeled dependencies among content words in the original input. LFs normalize certain syntactic alternations (e.g. active/passive) and resolve both intrasentential anaphora and long-distance dependencies.

Over the past two years we have been exploring the use of MEG LFs as a means of

improving IR precision. This work, which is embodied in a natural language query feature in the Microsoft Encarta 99 encyclopedia, augments a traditional keyword document index with a second index that contains linguistically-informed terms. Two types of terms are stored in this linguistic index:

1. *LF triples.* These are subgraphs extracted from the LF. Each triple has the form *word$_1$-relation-word$_2$*, describing a dependency relation between two content words. For example, for the sentence *Abraham Lincoln, the president, was assassinated by John Wilkes Booth*, we extract the following LF triples:[1]

assassinate—LSubj—John_Wilkes_Booth
assassinate—LObj—Abraham_Lincoln
Abraham_Lincoln—Equiv—president

2. *Subject terms.* These are terms that indicate which words served as the grammatical head of a surface syntactic subject in the document, for example:

Subject: Abraham_Lincoln

This linguistic index is used to postfilter the output of a conventional statistical search algorithm. An input natural language query is first submitted to the statistical search algorithm as a set of content words, resulting in a ranked set of documents. This ranked set is then re-ranked by attempting to find overlap between the set of linguistic terms stored for each of these documents and corresponding linguistic terms determined by processing the query in MEG. Documents that contain linguistic matches are heuristically ranked according to the nature of the match. Documents that fail to match do not receive a rank, and are typically not displayed to the user. The process of building a secondary linguistic index and matching terms from the query is referred to as natural language matching (NLM) in the discussion below. NLM has been used to filter documents retrieved by several different search technologies operating on different genres of text.

Since NLM was intended for use in consumer products, it was important to minimize index size. We needed an algorithm that would enable us to achieve reductions in index size without adversely affecting precision and recall. At the time when we were conducting these experiments, there did not exist any sufficiently large publicly available corpora of questions and relevant documents for the two genres of interest to us: the world wide web and encyclopedia text. We therefore gathered queries and documents for a web sample (section 3.2) and Encarta 99 (section 3.3), and had non-linguists perform double-blind evaluations of relevance.

Three implementation-specific aspects of the NLM index should be noted. First, in order to limit index size, duplicate instances of a term occurring in the same document are stored only once. Second, because of the particular compression scheme used to build the index, all terms require the same number of bits for storage, regardless of the length or number of words they contain. Third, the top ten percent of the NLM terms were suppressed, by analogy with stop words in conventional indexing schemes. Such high frequency terms tended not to be good predictors of document relevance.

3 Experiments

We conducted experiments in which we eliminated terms from the NLM index, and then measured precision and recall. The experiments were performed on two test corpora: web pages returned by the Alta Vista search service (section 3.2) and articles from the Encarta electronic encyclopedia (section 3.3).

3.1 The kinds of subordinate clauses

In order to test the hypothesis that information contained in subordinate clauses is less useful for IR than matrix clause information, we modified the indexing algorithm so that it eliminated terms that occurred in certain kinds of subordinate clauses. We experimented with the following clause types:

[1] *LSubj* denotes a logical subject, *LObj* a logical object and *Equiv* an equivalence relation.

Abbreviated Clause (ABBCL)
Until further indicated, lunch will be served at 1 p.m.

Complement Clause (COMPCL)
I told the telemarketer that you weren't home.

Adverbial Clause (ADVCL)
After John went home, he ate dinner.

Infinitival Clause (INFCL)
John decided to go home.

Relative Clause (RELCL)
I saw the man, who was wearing a green hat.

Present Participial Clause (PRPRTCL)
Napoleon attacked the fleet, completely destroying it.

In the experiments described below, terms were eliminated from documents during indexing. However, terms were never eliminated from the queries.

3.2 Alta Vista experiments

We gathered 120 natural language queries from colleagues for submission to Alta Vista.[2] The queries averaged 3.7 content words, with a standard deviation of 1.7.[3] The following are illustrative of the queries submitted:

Are there any air-conditioned hotels in Bali?
Has anyone ported Eliza to Win95?
What are the current weather conditions at Steven's Pass?
What makes a cat purr?
Where is Xian?
When will the next non-rerun showing of Star Trek air?

We examined the first thirty documents returned by Alta Vista (or fewer documents for queries that did not return at least thirty documents). This document set comprised 3,440 documents. Since we were not able to determine what percentage of the web Alta Vista accounted for, it was not possible to calculate the recall of this document set. In the discussion below, we calculate recall as a percentage of the relevant documents returned by Alta Vista. Precision and recall are averaged across all queries submitted to Alta Vista. The documents returned by Alta Vista were indexed using NLM (section 2) and filtered to retain only documents that contained matches.

Table 1 contrasts the baseline NLM figures (indexing based on terms in all clauses) with the results of eliminating from the documents all terms that occurred in subordinate clauses.

To measure the trade-off between precision and recall, we calculated the F-measure (Van Rijsbergen 1980), defined as $F = \frac{(\beta^2 + 1.0)PR}{\beta^2 P + R}$, where P is precision, R is recall and β is the relative weight assigned to precision and recall (for these experiments, $\beta = 1$).

As Table 1 shows, by eliminating terms from all subordinate clauses in the documents, the NLM index size was reduced by 31.4% with only a minor impact (-0.82%) on F-measure. Given unique indexing of terms per document, and a constant size per term (section 2), we can deduce that 31.4% of the terms in the NLM index occurred *only* in subordinate clauses. Had they occurred even once in a main clause, they would not have been removed from the index.

We ran two comparison experiments. In the first comparison, we deleted one third of all terms as they were produced. Table 2 gives the average results of three runs of this experiment. In each run, a different set of one third of the terms was deleted. Although fewer terms were omitted (28.8%[4] versus 31.4% when all terms in

[2] Alta Vista's main search page (http://altavista.com) encourages users to submit natural language queries.

[3] Words like "know" and "find", which are common in natural language queries, are included in these counts.

[4] Terms eliminated from a subordinate clause in one sentence might persist in the index if they occurred in the main clause of another sentence in the same document, hence a reduction of slightly less than 33.3%.

subordinate clauses were eliminated) the detrimental effect on F-measure was 5.3 times greater than when terms occuring in subordinate clauses were deleted.

Table 1 Alta Vista: Effects of eliminating subordinate clauses

Algorithm	Precision	Recall	F	% Change in F[5]	% Change in index size
Baseline NLM	34.3	43.2	38.24	0.00	0.0
Subordinate clauses	35.9	40.2	37.93	-0.82	-31.4

Table 2 Alta Vista: Average effect of eliminating one third of terms

Precision	Recall	F	% Change in F	% Change in index size
36.9	36.4	36.65	-4.34	-28.8

In the second comparison experiment, we tested the converse of the operation described in the discussion of Table 1 above: we eliminated all search terms from the main clauses of documents, leaving only search terms that occurred in subordinate clauses. Table 3 shows the dramatic effect of this operation: as we expected, the index size was greatly reduced (by 73.8%). However, F-measure was seriously affected, by more than two thirds, or –68.99%. The effect on F-measure is primarily due to the severe impact on recall, which fell from a tolerable baseline of 43.2% to an unacceptable 7.5%. Comparing the reduction in index size to the reduction when subordinate clause information was eliminated (73.8% versus 31.4%, a factor of approximately 2:1) to the reduction in F-measure (-68.99 versus -0.82, a factor of approximately 84:1), it is clear that the impact on F-measure from eliminating terms in main clauses is disproportionate to the reduction in index size.

Table 3 Alta Vista: Effect of eliminating main clauses

Precision	Recall	F	% Change in F	% Change in index size
28.3	7.5	11.86	-68.99	-73.8

Table 4 isolates the effects of deleting each kind of subordinate clause. Most remarkable is the fact that eliminating terms that only occur in relative clauses (RELCL) yields a 7.3% reduction in index size while actually improving F-measure. Also worthy of special note is the fact that two kinds of subordinate clauses can be eliminated with no perceptible effect on F-measure: eliminating complement clauses (COMPCL), yields a reduction in index size of 7.4%, and eliminating present participial clauses (PRPRTCL) yields a reduction in index size of 4.2%.

[5] F is calculated from the underlying figures, to minimise the effects of rounding errors.

Table 4 Alta Vista: Effect of eliminating different kinds of subordinate clauses

Algorithm	Precision	Recall	F	% Change in F	% Change in index size
Baseline NLM	34.3	43.2	38.24	0.00	0.0
ADVCL	34.6	42.1	37.98	-0.67	-7.0
ABBCL	34.3	43.2	38.24	0.00	-0.3
INFCL	34.8	42.1	38.10	-0.36	-11.8
RELCL	34.9	42.6	38.37	0.33	-7.3
COMPCL	34.5	42.9	38.24	0.00	-7.4
PRPRTCL	34.5	42.9	38.24	0.01	-4.2

Because of interactions among the different clause types, the effects illustrated in Table 4 are not additive. For example, an infinitival clause (INFCL) may contain a noun phrase with an embedded relative clause (RELCL). Elimination of all terms in the infinitival clause would therefore also lead to elimination of terms in the relative clause.

3.3 Encarta experiments

We gathered 348 queries from middle-school students for submission to Encarta, an electronic encyclopedia. The queries averaged 3.4 content words, with a standard deviation of 1.4. The following are illustrative of the queries submitted:

How many people live in Nebraska?
How many valence electrons does sodium have?
I need to know where hyenas live.
In what event is Amy VanDyken the closest to the world record in swimming?
What color is a giraffe's tongue?
What is the life-expectancy of an elephant?

We indexed the text of the Encarta articles, approximately 33,000 files containing approximately 576,000 sentences, using a simple statistical indexing engine. We then submitted each query and gathered the first thirty ranked documents, for a total of 5,218 documents. We constructed an NLM index for the documents returned and, in a second pass, filtered documents using NLM. In the discussion below, recall is calculated as a percentage of the relevant documents that the statistical search returned.

Table 5 compares the baseline NLM accuracy (indexing all terms) to the accuracy of eliminating terms that occurred in subordinate clauses. The reduction in index size (29.0%) is comparable to the reduction observed in the Alta Vista experiment (31.4%). However, the effect on F-measure of eliminating terms from subordinate clauses is more marked (-4.91%) than in the Alta Vista experiment (-0.82%).

Table 5 Encarta: Effects of eliminating subordinate clauses

Algorithm	Precision	Recall	F	% Change in F	% Change in index size
Baseline NLM	39.2	29.0	33.34	0.00	0.0
Subordinate clauses	41.1	25.9	31.78	-4.91	-29.0

The impact on F-measure is still substantially less than the average of three runs during which arbitrary non-overlapping thirds of the terms were eliminated, as illustrated in

Table 6. This arbitrary deletion of terms results in an 11.57% reduction in F-measure compared to the baseline, approximately 2.4 times greater than the impact of eliminating material in subordinate clauses.

Table 6 Encarta: Effects of eliminating one third of terms

Precision	Recall	F	% Change in F	% Change in index size
40.2	23.8	29.88	-11.57	-29.5

As Table 7 shows, eliminating terms from main clauses and retaining information in subordinate clauses has a profound effect on recall for the Encarta corpus. As with the Alta Vista experiment (section 3.2), it is instructive to compare the results in Table 7 to the results obtained when terms in subordinate clauses were deleted (Table 5). Approximately 2.7 times as many terms were eliminated from the index, yet the effect on F-measure is almost thirteen times worse.

Table 7 Encarta: Effect of eliminating main clauses

Precision	Recall	F	% Change in F	% Change in index size
40.9	7.4	12.53	-62.41	-77.1

Table 8 isolates the effects for Encarta of eliminating terms from each kind of subordinate clause. It is interesting to compare the reduction in index size and the relative change in F-measure for Encarta, a relatively homogeneous corpus of academic articles, to the heterogeneous web sample of section 3.2. For both corpora, eliminating terms that only occur in abbreviated clauses (ABBCL) or present participial clauses (PRPRTCL) results in modest reductions in index size without negatively affecting F-measure. Eliminating terms from adverbial clauses (ADVCL) or infinitival clauses (INFCL) also produces a similar effects on the two corpora: a reduction in index size with a modest (less than 1%) reduction in F-measure. Relative clauses (RELCL) and complement clauses (COMPCL), however, behave differently across the two corpora. In both cases, the effects on F-measure are positive for web documents and negative for Encarta articles. The negative impact of the elimination of material from relative clauses in Encarta can perhaps be attributed to the pervasive use of non-restrictive relative clauses in the definitional encyclopedia text, as illustrated by the underlined sections of the following examples:

Sargon II (ruled 722-705 BC), who followed Tiglath-pileser's successor, Shalmaneser V (ruled 727-722 BC), to the throne, extended Assyrian domination in all directions, from southern Anatolia to the Persian Gulf.

Amaral, Tarsila do (1886-1973), Brazilian painter whose works were instrumental in the development of modernist painting in Brazil.

After the so-called Boston Tea Party in 1773, when Bostonians destroyed tea belonging to the East India Company, Parliament enacted four measures as an example to the other rebellious colonies.

Another peculiar characteristic of the Encarta corpus, namely the pervasive use of

complement taking nominal expressions such as *the belief that* and *the fact that*, possibly explains the negative impact of the elimination of complement clause material in Table 8.

Table 8 Encarta: Effect of eliminating different kinds of subordinate clauses

Algorithm	Precision	Recall	F	% Change in F	% Change in index size
Baseline NLM	39.2	29.0	33.34	0.00	0.0
ADVCL	39.9	28.4	33.18	-0.47	-5.8
ABBCL	39.6	29.0	33.48	0.43	-0.4
INFCL	40.0	28.3	33.15	-0.57	-9.2
RELCL	39.7	28.2	32.98	-1.10	-9.5
COMPCL	38.9	28.3	32.76	-1.75	-3.3
PRPRTCL	39.8	29.0	33.55	0.64	-5.5

4 Discussion

Although the results presented in section 3 are compelling, it may be possible to refine the identification of clauses from which index terms can be eliminated. In particular, complement clauses subordinate to speech act verbs would appear from failure analysis to warrant special attention. For example, in the following sentence our linguistic intuitions suggest that the content of the complement clause is more informative than the attribution to a speaker in the main clause: *John said that the President would not resign in disgrace.* Of course, more fine-grained distinctions of this type can only be made given sufficiently rich linguistic analyses as input. Another compelling topic for future research would be the impact of less sophisticated analyses to identify various kinds of subordinate clauses.

The terms eliminated in the experiments presented in this paper were linguistic in nature. However, we would expect similar results if conventional word-based terms were eliminated in similar fashion. In future research, we intend to experiment with eliminating terms from a conventional statistical engine, combining this technique with the standard method of eliminating high frequency index terms.. Rather than eliminating terms from an index, it may also prove fruitful to investigate weighting terms according to the kind of clause in which they occur.

5 Conclusions

We have demonstrated that, as implicitly predicted by RST, index terms may be eliminated from certain kinds of subordinate clauses without substantially affecting precision or recall. Rather than using NLP to generate more index terms, we have found tremendous gains from systematically eliminating terms. The exact severity of the impact on precision and recall that results from eliminating terms varies by genre. In all cases, however, the systematic elimination of subordinate clause material is substantially better than arbitrary deletion of index terms or the deletion of index terms that occur only in main clauses.

Future research shall attempt to refine the analysis of the kinds of subordinate clauses from which index terms can be omitted, and to integrate these findings with conventional statistical IR algorithms.

Acknowledgements

Our thanks go to Lisa Braden-Harder, Susan Dumais, Raman Chandrasekar, Eric Ringger, Monica Corston-Oliver, Lucy Vanderwende and the three anonymous reviewers for their help and comments on an earlier draft of this paper and to Jing Lou for assistance in configuring a test environment.

References

Arampatzis, A. T., T. Tsoris, C. H. A. Koster, T. P. Van Der Weide. (1998) "Phrase-based information retrieval", *Information Processing and Management* 34:693-707.

Corston-Oliver, S. H. (1998) *Computing Representations of the Structure of Written Discourse.* Ph.D. dissertation. University of California, Santa Barbara.

Fagan, J. L. (1988) *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-syntactic Methods.* Ph.D. dissertation. Cornell University.

Heidorn, G. (1999) "Intelligent writing assistance." To appear in Dale, R., H. Moisl and H. Somers (eds.), *A Handbook of Natural Language Processing Techniques.* Marcel Dekker.

Katz, B. (1997) "Annotating the World Wide Web Using Natural Language." *Proceedings of RIAO 97, Computer-assisted Information Search on Internet*, McGill University, Quebec, Canada, 25-27 June 1997. Vol. 1:136-155.

Mann, W. C. and Thompson, S. A. (1986) "Relational Propositions in Discourse." *Discourse Processes* 9:57-90.

Mann, W. C. and Thompson, S. A. (1988) "Rhetorical Structure Theory: Toward a functional theory of text organization." *Text* 8:243-281.

Matthiessen, C. and Thompson, S. A. (1988) "The structure of discourse and 'subordination'." In Haiman, J. and S. A. Thompson, (eds.). 1988. *Clause Combining in Grammar and Discourse.* John Benjamins: Amsterdam and Philadelphia. 275-329.

Strzalkowski, T. G. Stein, G. B. Wise, J. Perez-Carball, P. Tapanainen, T. Jarvinent, A. Voutilainen, J. Karlgren. (1997) *Natural Language Information Retrieval: TREC-7 Report.*

Van Rijsbergen, C. J. (1980) *Information Retrieval.* Butterworths: London and Boston.