The Use of an Index Term Corpus for the Development of an Automatic Indexer

Timo Lahtinen

University of Helsinki, Department of General Linguistics

Abstract

This paper presents a linguistic approach to the problem of extracting index terms automatically. Promising results were obtained by developing a linguistic indexer which weights the expressions of a text according to their index-term-likeness. The typical linguistic features of index terms were explored using a linguistically analysed text collection in which the index terms are manually marked up. This text collection is referred to as an *index term corpus*. Specific features of the index terms provided the basis for a linguistic term-weighting scheme, which was then combined with a frequency-based term-weighting scheme. The use of this kind of index term corpus as training material is a new method for developing an automatic indexer.

1 Introduction

The main point of this paper is to illustrate the process of developing an automatic indexer based on a linguistic analysis of an index term corpus. In the index term corpus *manually* generated index terms are marked up by tags, and their linguistic features are explored.

An index term is an expression which contains a considerable amount of information about the contents of a text; for example, an index in a book consists of terms that refer to key content included in the book, such as concepts, persons, events. In information retrieval systems, an index language is the language that describes the documents and queries, and index terms (or descriptors or keywords) are the elements of the index language. Indexing may be done automatically or by human indexers, and index terms may be expressions derived from the text or expressions defined independently.

In information retrieval systems, index terms are usually weighted according to their importance for describing documents. Typically the weighting schemes are based on word frequencies across the document collection. In experiments using natural language processing techniques to improve retrieval performance, the role of linguistic analysis is often restricted to discovery of multi-word phrases for indexing. These terms are then weighted by some frequency-based weighting technique. The weighting scheme in this paper, however, combines evidence derived from word distributions with evidence derived from linguistic analysis.

2 The index term corpus

2.1 Texts and indexes

The index term corpus in this study consisted of five texts that dealt with sociology and philosophy. Four were essays[1] and the fifth was a longer document[2]. All texts had manually generated indexes. A research aide identified and marked up the index terms for each document page using the document index. The corpus was then analysed linguistically and divided into two parts: *a training corpus*, consisting of two essays and 57 pages from the long text, and *a test corpus*, consisting of the remaining two essays and 16 pages. The features of index terms were explored using the training corpus, and the test corpus was used to test whether the results could be generalized beyond the context of the training corpus. In addition, there were fifteen texts with no index term mark-up, to supplement the index term corpus when applying frequency-based term weighting schemes.[3] In short, the total corpus consisted of 629,961 words:

- a training corpus of 38,136 words with index term mark-up
- a test corpus of 17,392 words with index term mark-up
- a corpus of 574,433 words with no index term mark-up

In the indexes, terms were usually simple noun phrases, but in the texts the content of the noun phrases were sometimes expressed by using verbs, adjectives, or even clauses. For instance, the index term `oppression of woman` was expressed in the text by the clause `women are oppressed`. In such cases, the research aide was instructed to mark up the closest equivalents. Sometimes index terms were nested, e.g. `dominant structure` was included in the index as two index terms, `dominant structure` and `structure`. In these cases, both index terms were marked up.

In the training corpus, a single word was marked up as an index term 2,145 times out of 38,136. A bigram was considered as an index term 1,183 times, and terms with more than two words had a frequency of 309. The most typical index term patterns found were simple noun phrases, for instance, `capitalism`, `biological determinism` and `methodology of philosophy`. Not surprisingly, almost all proper nouns in the text were included in the index.

2.2 Linguistic annotation

The linguistic annotation of the corpus was done with a robust rule-based dependency parser, the Conexor Functional Dependency Grammar (FDG, cf. Tapanainen and Järvinen, 1997 [4]), which is related to the Constraint Grammar framework

[1] Together, these essays form sample ECV from the British National Corpus.

[2] A 73-page document taken from the Bank of English.

[3] The texts were again taken from the British National Corpus (A6S, APD, CGF, CM6, CMN, CMR, CRF, EDH, F9K, F9V, FAC, GV5, GVA, H9F and J2K).

[4] For a demo, see: http://www.conexor.fi/analysers.html#testing

(Karlsson *et al.*, 1995). The dependency parser creates links between the elements of the sentence in addition to the shallow representation, similar to English Constraint Grammar (ENGCG) (Voutilainen, 1994; Järvinen, 1994). The parser also applies an ENGTWOL-style lexicon (Heikkilä, 1995; Koskenniemi, 1983), and morphological disambiguator designed by Voutilainen (1995).

The following example is from the training corpus:

```
"<Marx>"
    "Marx" <Proper> N NOM SG @SUBJ subj:>2 </INDEX>
"<suggested>"
    "suggest" V PAST VFIN @+FMAINV #2 main:>0
```

In the first place, each word is annotated with a base form, which is a useful feature for counting word frequencies. Then, the tag list[5] carries information about the linguistic features of the individual words, e.g.

- `<Proper>` proper noun
- `N` noun
- `NOM` nominative case
- `@+FMAINV` finite main predicator

and about the dependency links between the words, e.g. `Marx` is the subject of the sentence (`subj:>2`), and it has a link to the main verb (`suggest #2`).

Furthermore, a research aide manually marked up all index terms by `</INDEX>`-tags, which were added to the automatically generated tag lists.

2.3 Linguistic structure of index terms

The combination of linguistic annotation and index term mark-up made it possible to examine the linguistic structure of index terms (Figure 1). We find index terms consisting of verbs (e.g. `understand`), adverbs (e.g. `historically`), adjectives (e.g. `empirical`) and even clauses (e.g. `women are oppressed`), but the great majority of index terms were noun phrases.

The most common pattern (`A-N`) consisted of an adjective as a premodifier and a noun as a head. Single common nouns (`N`) comprised the next largest group of index terms. Proper nouns (`prop`) included all of the proper noun terms of various lengths. The pattern of two successive nouns (`N-N`) contained a few genitive constructions, such as `women's oppression`, and a number of compounds, e.g. `mass media`. Two successive premodifiers (`attr:>`) in the `a-a-N`-pattern were either nouns or adjectives, for instance, `Marx's scientific socialism` and `oppressive social structures`. Genitive constructions using *of*-preposition (`of`) included 92 different index terms of various lengths, e.g. `oppression of women` and `structural and historically specific nature of capitalism`. 92 index terms contained postmodification with the

[5]For ENGTWOL and ENGCG tag descriptions, cf., Voutilainen *et al.*, 1992.

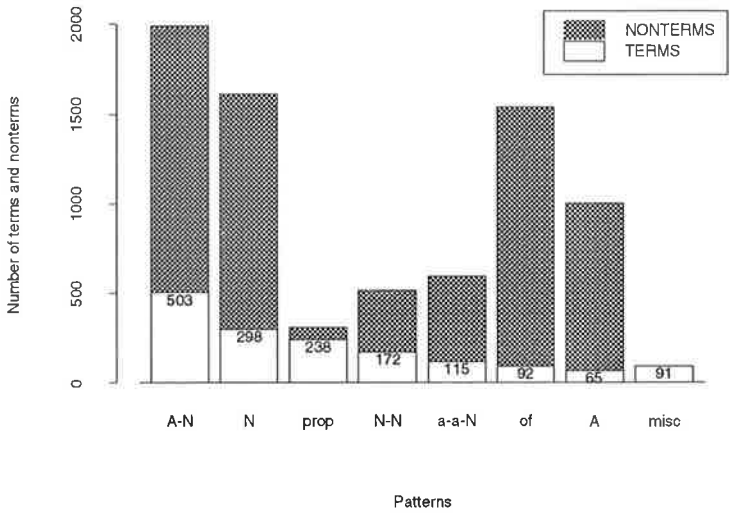

Figure 1: Different term patterns and their frequencies in the training corpus.

preposition *of*. The only major non-NP group was formed by 65 index terms consisting of single adjectives (`A`). Only 91 index terms (`misc`) did not fall into any one of the seven above-mentioned categories.

The length of the most common index term patterns varied from one to three words. Only 95 index terms were longer (50 of which showed *of*-constructions), and some of these had only very few representatives. For example, three patterns consisted of five words each, but the training corpus included only 15 such terms in total (e.g. `surface of oppressive structural relationships`). Nine index terms were even longer than five words.

3 Frequency-based term weighting

3.1 Popular existing methods

It is usually attempted to identify potential index terms on the basis of their frequency in the documents. It is then the medium-frequency words which are considered to be the most content-bearing words and the best potential index terms. The most frequent words (e.g. *the*, *of*, *and*), are the least content-bearing and not appropriate as index terms, while the least-frequent words are also considered to be as

poor index terms, since they may only show random noise. However, the elimination of high-frequency and low-frequency words by using absolute word frequency measures may produce losses in recall. This suggests that weighting techniques based on relative frequency measures should be used. If a word occurs frequently in some documents, but has a low overall collection frequency, it is likely to be an appropriate index term. A typical weighting scheme is TF*IDF (van Rijsbergen, 1979; Salton and McGill, 1983), where TF is an abbreviation of "Term Frequency", and IDF is an abbreviation of "Inverse Document Frequency". Term frequency is the number of times a particular term occurs in a given document. Inverse document frequency is a measure of how often a particular term appears across the document collection. TF*IDF may be defined as

$$TF * IDF = TF * log \frac{Number\ of\ documents\ in\ a\ collection}{Number\ of\ documents\ containing\ the\ term}$$

Thus, words occurring in many documents have a low IDF and words unique to a document have a high IDF. The most appropriate index terms for a given document are those words with a high IDF and also with a high frequency in that particular document. In our experiments, we based the TF*IDF-weights on the base form of each word, as provided by the linguistic analysis, and use all twenty documents as data sources for the IDF-values.

In this project, the idea of the TF*IDF-formula was applied to another formula as well, referred to as SF*IPF. "Stem Frequency" (SF) is the number of times a particular stem occurs in a given document. "Inverse Paragraph Frequency" (IPF) is a measure of how many paragraphs contain the stem. SF*IPF may be defined as:

$$SF * IPF = SF * log \frac{Number\ of\ paragraphs\ in\ a\ document}{Number\ of\ paragraphs\ containing\ the\ stem}$$

Thus, a stem occurring in many paragraphs has a low IPF and a stem occurring frequently in a small number of paragraphs has a high SF*IPF. Note that SF*IPF does not try to find index terms for each separate paragraph but for a whole document, just like TF*IDF. For this reason SF is defined as 'the number of times a stem occurs in a *document*' instead of 'the number of times a stem occurs in a *paragraph*'. The stemming algorithm is based on the use of a list of suffixes and the removal of the longest word endings matching any suffix on the list, for instance:

WORD	STEM	WORD-FREQ	STEM-FREQ
ethnography	ethnograph	27	59
ethnographic	ethnograph	23	59
ethnographer	ethnograph	9	59

This means that `ethnography`, `ethnographic`, and `ethnographer` have the same SF*IPF-value. The number of different words in the 38,136 word training corpus was 3,587 and the number of different stems was 3,146. The purpose of the SF*IPF-formula is to sum up the frequencies of the different variations of an index

term so that the weight of the term will be higher. No stem matrix is constructed; the formula just uses stem frequencies instead of word frequencies in weighting the words.

3.2 Frequency-based weighting in this study

In principle, the frequency-based methods can be applied to both single-word and multi-word terms. In our study, however, it was difficult to apply these methods to multi-word terms as well. The reason is that, while the average frequency of single-word terms is 9.519, the average frequency of multi-word terms is only 1.833, and over 76% (711/940) of the multi-word terms occurred only once in the training corpus. As a result the frequency-based term weighting was restricted to single-word terms in this study.

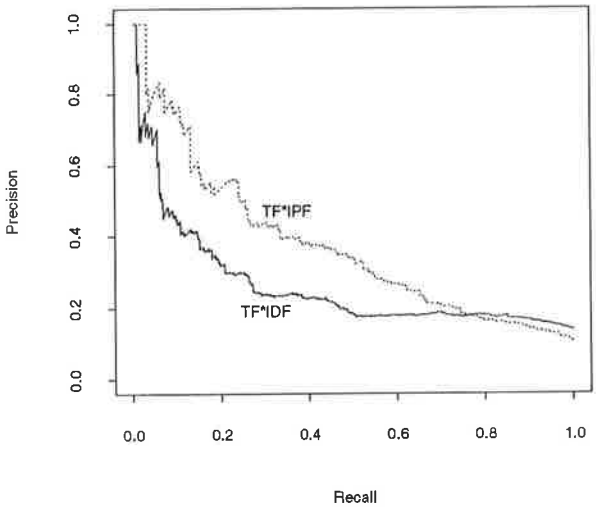

Figure 2: Term retrieval. Evaluation of TF*IDF and TF*IPF (test corpus).

The actual weighting scheme applied in this study was a combination of the two above-mentioned formulas: TF*IPF=(TF*IDF)*(SF*IPF). This combined formula was used since TF*IPF proved to predict index-term-likeness more reliably than TF*IDF or SF*IPF alone. Figure 2 indicates the advantage of the TF*IPF-weights over the TF*IDF-weights. Here, the words of the test corpus are ranked by the TF*IDF-weights and by the TF*IPF-weights. The evaluation of the ranked

word lists is illustrated by recall-precision curves, with the points representing the level of precision (the number of found terms divided by the number of scanned words) at different recall percentages (the number of found terms divided by the total number of terms).

TF*IPF-values were calculated separately for the three documents of the test corpus. Those words that were present in more than one document thus had more than one TF*IPF-value. In the TF*IPF-matrix of the test corpus, each word could occur only once, and the highest TF*IPF-value of a word was chosen to be the value of the TF*IPF-variable, as shown in the following example:

```
WORD                TF*IPF-variable
--------------------------------
abandon                 0.254
ability                 0.324
able                    0.296
abortion-decision       0.477
```

To sum up, the TF*IDF-, SF*IPF-, and TF*IPF-weights were all calculated for the words of a *document* instead of the words of a paragraph or the words of a corpus. Once the TF*IPF-weights were calculated for the words of different documents, the highest weight of each word was chosen to represent the value of the TF*IPF-variable in the TF*IPF-matrix of the test corpus.

4 Linguistic term weighting

4.1 Use of the training corpus

With the help of the linguistic analysis of the index term corpus, it becomes possible to try to identify index terms on the basis of their linguistic properties. A simple way to explore the typical features of index terms is to see how often each tag is included in the tag list of index terms. For example, the frequency of `N`-tag was 10,111 in the training corpus, and it appeared 1,987 times in the tag list of a single-word term. The training corpus gave an estimated index term probability of 0.197 (1,987/10,111) to the `N`-tag, a probability of 0.755 to the `<Proper>`-tag, a probability of 0.145 to the `subj:>`-tag (subject of the sentence), a probability of 0.082 to the `obj:>`-tag (object of the sentence), and so on. The probabilities for the different tags are obviously not independent, so they were calculated for all of the relevant *tag combinations*, i.e. for the combinations that distinguish index terms from non-terms in the most appropriate way. For instance, the tag combinations of the words `Marx` and `suggested` have the following index term probabilities:

```
"<Marx>"
    "Marx" <Proper> N NOM SG @SUBJ subj:>2 </INDEX> PROBABILITY:0.985
"<suggested>"
    "suggest" V PAST VFIN @+FMAINV #2 main:>0 PROBABILITY:0.005
```

Multi-word terms were studied in the same way as single-word terms. The patterns of index terms were explored by using the training corpus, and the index term

probabilities of the different tag combinations were calculated. The noun phrase `autonomous person`, for instance, is assigned the index term probability of 0.172 (40/233) in the following context:

```
"<An>"
    "an" <*> <Indef> DET CENTRAL ART SG @DN> det:>3
"<autonomous>"
    "autonomous" A ABS @A> attr:>3 <INDEX>
"<person>"
    "person" N NOM SG @SUBJ #3 subj:>4 </INDEX>
"<is>"
    "be" V PRES SG3 VFIN @+FMAINV #4 main:>0
```

The first word is an adjective (`A`) and a premodifier (`attr:>`), and the head is a noun (`N`) and a subject (`subj:>`). In the training corpus, this tag combination occurred 233 times, and of these, it was a tag combination of an index term 40 times.

Altogether, 85 different tag combinations were considered as relevant term patterns, and index term probabilities were calculated for these combinations based on the training corpus. Matrices of DEP-values were constructed for the multi-word terms in the same way as for the single-word terms.

4.2 Identification of single-word terms

The power of probabilities to predict which words are index terms was then tested by using the test corpus. Each word of the test corpus was given a certain index term probability based on the probabilities calculated from the training corpus. These probabilities were the values of the DEP-variable in a matrix in which the words were observations. The following are examples:

```
WORD                DEP-variable
--------------------------------
abandon               0.019
ability               0.609
able                  0.276
abortion-decision     0.380
```

Each word occurred in the matrix only once, and the value of the DEP-variable represented the highest probability of the word. Thus, if `structure` had the probability of 0.005 as a verb, and the maximum probability of 0.294 as a noun in a certain syntactic position, the DEP-value of `structure` was 0.294 in the matrix.

Figure 3 shows the recall-precision curves for the linguistic weighting method. It suggests that the probabilities calculated from the training corpus predicted the index-term-likeness of the words of the test corpus rather well. The recall-precision curve of DEP was also clearly better than the recall-precision curve of the generally used TF*IDF.

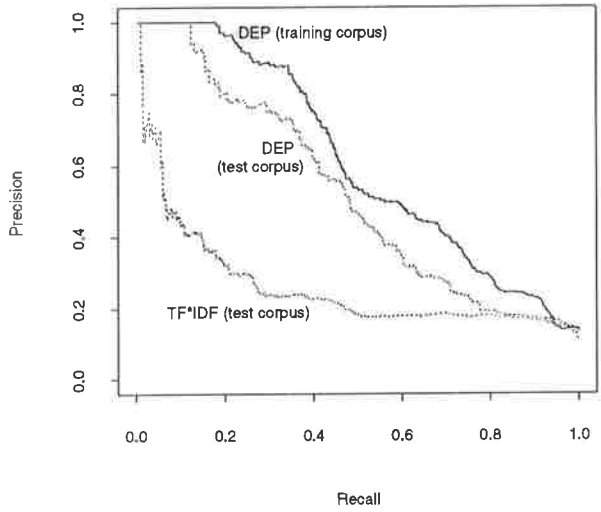

Figure 3: Evaluation of the weighting schemes.

4.3 Identification of multi-word terms

The identification of multi-word terms was satisfactory as well. The following examples from the ranking list give an idea of the quality of multi-word term weighting. Every tenth term candidate is extracted from the beginning of the ranking list. A plus sign (+) refers to a term, and a minus sign (-) to a non-term; the DEP-column contains the DEP-weight of the candidate, and the FREQ-column contains the frequency of the candidate:

TERM CANDIDATE (test corpus)	DEP	FREQ	RANK
+ working-class life	1.000	1	1
+ Sylvia Plath	1.000	1	11
+ Vauxhall Motors	0.868	1	21
+ John Rawls	0.868	1	31
+ Alison Assiter	0.868	1	41
+ working class	0.750	23	51
- standard quantitative social research process	0.500	1	61
- way man	0.455	2	71
- pupil misbehaviour	0.455	1	81
+ male desire	0.455	1	91

The index terms were ranked significantly higher than non-terms (Mann Whitney's U, $p > 0.95$). The 50 highest ranked term candidates included only four non-terms, and the next 50 terms included 34 non-terms. The 786 lowest ranked term candidates were all non-terms. The 71st candidate, `way man`, had the context `the way men typically see it`. This demonstrated a typical, albeit rare, mistake of the dependency parser: `way` was incorrectly analysed as a premodifier of `men`. It is not a trivial task for a parser to analyse two successive nouns correctly.

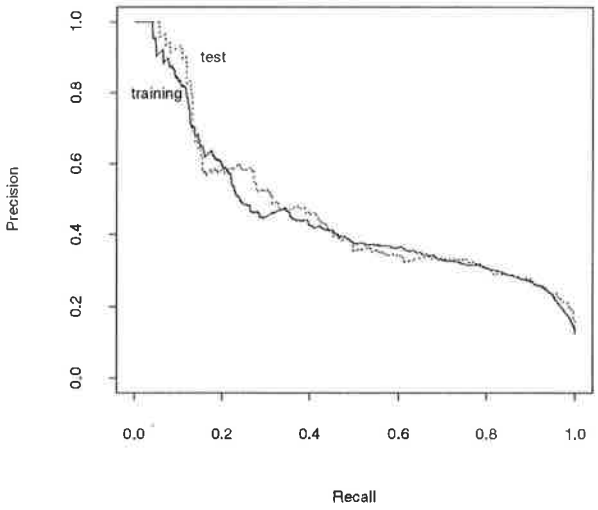

Figure 4: Multi-word terms.

The precision-recall curves (Figure 4) again suggest that the probabilities calculated from the training corpus predicted the index-term-likeness of the term candidates in the test corpus reasonably well. In fact, the recall-precision curve of the test corpus is even slightly better than the curve of the training corpus. One reason for this is that the training corpus contained more low-weight terms than the test corpus. The lowest ranked index term had a weight of 0.002 in the training corpus and a weight of 0.009 in the test corpus. Thus, in the test corpus, 100% recall was reached faster, and precision percentages were also higher at low recall percentages. The recall-precision curve of the single-word DEP-weights (Figure 3) is somewhat better than the curve of the multi-word DEP-weights. The most important reason for this is that index terms of more than two words were infrequent. For instance, the training corpus included only 15 terms of five words, and the test corpus only

four such terms. At the same time, 393 non-terms in the training corpus and 167 non-terms in the test corpus represented these three term patterns of five words. As a rule, long term patterns tended to have lower weights than short patterns.

5 Combining the linguistic and the frequency-based weighting schemes

5.1 Combination method

The frequency-based and linguistic weighting schemes presented above each have shown their merits in the prediction of the index-term-likeness of words. Furthermore, given their totally different starting-points, the two often assign high rank to different words. For example, proper nouns are always ranked highly by DEP-weighting, even if they occur only once in a document, whereas TF*IPF-weighting assigns a low rank to words that occur only once. On the other hand, verbs are always ranked low by DEP-weighting, whereas TF*IPF-weighting may well rank a verb high, depending on its distribution. This indicates that the results might be further improved by combining the methods.

An essential question at this point is how to combine the strengths of the linguistic and the frequency-based weighting approaches. To answer this question, the training corpus was used to explore a suitable method for doing this, and the method was tested by using the test corpus. The words observed in the training corpus were placed in a matrix, together with their weights (the variables DEP and TF*IPF), and the data in the matrix was then processed further by statistical methods. The idea was to partition the data in an appropriate way in order to optimize the result of combining the weighting schemes.

A regression tree model was used, which splits the data recursively into two parts until nodes are either homogeneous or the data is too sparse (Chambers and Hastie, 1991). Three variables were used here: TERM was the response variable that indicated whether the word was an index term. DEP and TF*IPF were the predictor variables, as shown in the example entries below:

WORD	DEP	TF*IPF	TERM
abandon	0.019	0.254	0
ability	0.609	0.324	0
able	0.276	0.296	0
abortion-decision	0.380	0.477	1

At any node, the predictor variable and the value of the predictor variable is selected which maximally distinguishes the response variable. Figure 5[6] presents a pruned tree that splits the data into six partitions. Pruning selects only the most important splits, and produces a rough classification of the data. In this case, six categories were allowed.

The tree illustrates, for example, that if the value of the DEP-variable was greater than 0.9285, the percentage of terms was 97.62% (the right-most leaf). If

[6]S-PLUS : Copyright (c) 1988, 1996 MathSoft, Inc. Version 3.4 Release 1.

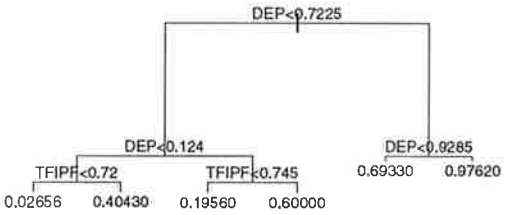

Figure 5: Division of candidates by a regression tree.

the value of DEP-variable was less than 0.124, and the value of TF*IPF-variable was less than 0.72, the percentage of terms was only 2.656% (the left-most leaf). `Abandon`, for example, belongs to this latter group (the DEP-value was 0.019, and the TF*IPF-value was 0.254). The second leaf from the left indicates that the tree model was able to partition the data into relevant groups: if the value of the DEP-variable was less than 0.124, and the value of the TF*IPF-variable was greater than 0.72, the percentage of terms was 40.43%. In other words, this group includes those index terms ranked high by the frequency-based weighting, and ranked low by the linguistic weighting. This partitioning of the data was then used in combining DEP and TF*IPF by a simple regression method.

The linear regression model calculates coefficients for variables by using the method of least squares to fit (Belsley, Kuh and Welsch, 1980). The variables were summed by the following formula:

$$DEP + (TF * IPF) = Intercept + Coefficient1 * DEP + Coefficient2 * (TF * IPF)$$

Separate *Intercept*- and *Coefficient*-values were calculated for each of the six groups of observations described above. For instance, the weight of `abandon` was then:

```
-0.0128 = -0.0742 + 0.6474*0.019 + 0.1935*0.254
```

The considerable differences between the coefficients of the groups indicate that partitioning the data was probably a relevant phase before calculating the coefficients.

This experiment did not include an investigation of whether the tree model and the linear regression model are the optimal methods for combining the linguistic

and frequency-based weighting; thus the results of the next section are to be considered as preliminary results.

5.2 Results of the combined method

The data of the test corpus consisted of 2,580 different words including 242 different single-word terms.[7] As a result of combining DEP and TF*IPF, the 35 highest ranked words of the test corpus were all terms. The first non-term was `Metamorphosis` (Kafka's Metamorphosis), and although it was not included in a book index, it could perhaps be considered as a potential index term.[8] The following sample drawn from the ranking list presents the 15 highest ranked and 5 lowest ranked words of the test corpus, as well as 5 words from the middle of the rankings (recall=50%, that is, half of the terms were ranked higher and the other half lower). A plus sign (+) refers to a term, and a minus sign (-) to a non-term:

	WORD (test corpus)	DEP+TF*IPF	RANK
+	love-making	1.0559	1
+	fantasy	1.0477	2
+	Willis	1.0283	3
+	Hegel	1.0283	4
+	Kohlberg	1.0256	5
+	Goldthorpe	1.0230	6
+	Rawls	1.0220	7
+	Gilligan	1.0207	8
+	Lockwood	1.0202	9
+	Kant	1.0175	10
+	Nozick	1.0162	11
+	worker	1.0114	12
+	capitalism	1.0098	13
+	culture	1.0087	14
+	Eisenstein	1.0076	15
.			
.			
-	upbringing	0.2716	221
+	Filmer	0.2703	222
-	universality	0.2702	223
-	employment	0.2702	224
+	Kantian	0.2699	225
.			
.			

[7] Since frequency-based weighting was not used for multi-word terms, the combined method also only considers single-word terms.

[8] Defining the content-bearing units of the text demands more or less subjective decisions, and a user of an index does not necessarily share the indexer's view. In any case, an index of a book represents an interpretation of the contents of the text. The recall and precision values of the test corpus were calculated by using the marked-up index terms of the books as the benchmark. Consequently, a number of highly ranked non-terms could in fact be appropriate index terms.

```
- single          -0.0241    2199
- serious         -0.0241    2200
- fall            -0.0247    2201
- contain         -0.0261    2202
- pass            -0.0267    2203
```

The index terms were ranked significantly higher than non-terms (Mann Whitney's U, $p > 0.95$). The ranking list contains only 2,203 words instead of the original 2,580 words, because the linguistic weighting method was able to eliminate obvious non-terms, such as pronouns, articles, quantifiers, negators, conjunctions and prepositions.

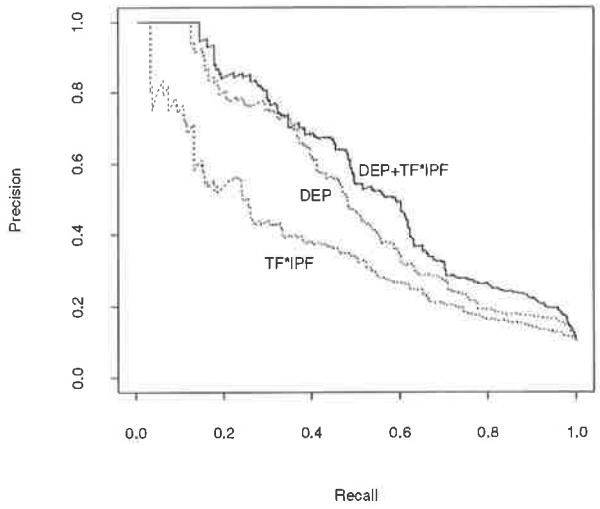

Figure 6: Evaluation of DEP+TF*IPF (test corpus).

Figure 6 presents the result of combining the linguistic and the frequency-based weighting approaches. The values of DEP (test corpus) and TF*IPF (test corpus) were multiplied by the coefficients calculated from the training corpus, and the products were summed. DEP+TF*IPF was found to predict index-term-likeness reasonably well: the recall-precision curve shows 57% precision at 50% recall (figure 6). The lowest ranked index term was the verb `transcend`, which had the form of `transcendence` in the book index. Precision would naturally be improved if the set of term candidates were restricted to nouns. Nouns, however, are not the only content-bearing elements of a text, although noun phrases do comprise the great majority of the index terms.

6 Conclusions

The results suggest that it is possible to define a number of typical features of index terms in order to develop an automatic indexer. In general, the index terms of the test corpus shared the features of the index terms of the training corpus. All texts of the corpus, however, represented the same genre, which must partly explain the promising results. If a text of a different genre were used as test material, the results would possibly not be as good. A robust indexing tool will require a large corpus of different texts as training material.

In this experiment, the frequency-based weighting schemes could not predict the index-term-likeness as reliably as the linguistic technique, but a larger corpus would probably improve its performance somewhat. The results support the assumption that integrating the linguistic and the frequency-based techniques would be a profitable approach to developing tools for information retrieval tasks. For instance, the index term `industrialism` occurred only once in the documents of the test corpus, and consequently, it was ranked low by the TF*IPF-weights. On the other hand, because of its tag list, it was ranked high by the DEP-weights. Another index term, `biological`, is an adjective, and so it was ranked low by the DEP-weights. However, because of its distribution, the word was ranked high by the TF*IPF-weights. In both cases, one weighting scheme overlooked an index term that was highly ranked by the other. If the weighting schemes are combined, the recall-precision curve can be improved, as the results of the previous section have indicated.

The DEP-weighting scheme has one remarkable advantage over the frequency-based weighting schemes. Once the probabilities of the tag combinations have been calculated, a sentence is a sufficient input for weighting the words, i.e., no document or document collection is needed. For instance, the words of the query `What role does Islam play in restricting women in Pakistan?` are weighted as follows:

```
Islam          0.985
Pakistan       0.766
woman          0.103
role           0.082
restrict       0.027
do             0.005
play           0.005
in             0.000
what           0.000
```

An automatic indexer weights the words of the documents and queries, and an information retrieval system uses the weights in retrieving the most relevant documents. Obviously, it is possible to use the multi-word terms in the same way, as well.

To sum up, this experiment in combining a linguistic weighting scheme with a modified version of the standard TF*IDF-weighting scheme has offered promising results. Since the index terms were explicitly marked up in the corpus, it proved

to be a relatively straightforward task to determine the basis for a simple linguistic weighting method, as well as to evaluate the performance of different weighting schemes. The dependency parser provides rich information on the linguistic features of index terms for the purpose of developing an automatic indexer, and it is possible to make this indexer more robust by constructing a larger index term corpus of a wide range of genres. Another subject of future research will be the evaluation of different techniques for combining the linguistic and frequency-based weighting schemes.

Acknowledgements

This research was done in the context of the graduate school "Linguistic Meaning and its Processing", and it was partially supported by the Research Unit for Multilingual Language Technology at the Department of General Linguistics at the University of Helsinki. I would like to thank all reviewers, and Hans van Halteren in particular, for their contribution and insightful comments on earlier drafts.

References

Belsley, D.A., E. Kuh, and R.E. Welsch (1980) *Regression Diagnostics*. Wiley, New York.

Chambers, J.M. and T.J. Hastie (Eds.) (1991) *Statistical Models in S*. Chapman & Hall.

Heikkilä, J. (1995) ENGTWOL English lexicon: solutions and problems. In Karlsson *et al.* (1995).

Järvinen, T. (1994) Annotating 200 Million Words: The Bank of English Project. In *COLING-94. The 15th International Conference on Computational Linguistics Proceedings*, volume 1, Kyoto, Japan, pp. 565-568.

Karlsson, F., A. Voutilainen, J. Heikkilä, and A. Anttila (Eds.) (1995) *Constraint Grammar: a language-independent system for parsing unrestricted text, volume 4 of Natural Language Processing*. Mouton de Gruyter, Berlin and New York.

Koskenniemi, K. (1983) Two-level morphology: a general computational model for word-form recognition and production. Publications No. 11. Department of General Linguistics, University of Helsinki.

van Rijsbergen, C.J. (1979) *Information Retrieval*. Second edition. Butterworth & Co Ltd, London.

Salton, G. and M.J. McGill (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Tapanainen, P. and T. Järvinen (1997) A non-projective dependency parser. In the Proceedings of the 5th Conference on Applied Natural Language Processing, Washington, D.C., April. ACL.

Voutilainen, A., J. Heikkilä, and A. Anttila (1992) Constraint Grammar of English. A Performance-Oriented Introduction. Publications No. 21, Department of General Linguistics, University of Helsinki.

Voutilainen, A. (1994) Designing a Parsing Grammar. Publications No. 22, Department of General Linguistics, University of Helsinki.
Voutilainen, A. (1995) Morphological disambiguation. In Karlsson *et al.* (1995).