

Evaluating Feature Selection Techniques on Semantic Likeness

Henrik Oxhammar

Stockholm University

henrik.oxhammar@ling.su.se

Abstract

In this paper, we describe a first in a series of experiments for determining the usefulness of standard feature selection techniques on the task of enlarging large semantic knowledge sources. This study measures and compares the performance of four techniques, including odds ratio, chi-square, and correlation coefficient. We also include our own procedure for detecting significant terms that we consider as a baseline technique. We compare lists of ranked terms extracted from a medical corpus (OHSUMED) to terms in a medical vocabulary (MeSH).

Results show that all four techniques tend to rank significant terms higher than less significant terms, although chi-square and correlation coefficient clearly outdo the other techniques on this test. When comparing the order of terms with their semantic relatedness to particular concepts in our gold standard, we notice that our baseline technique suggests orderings of terms that conform more closely to the conceptual relations in the vocabulary.

1 Introduction

Controlled vocabularies¹ are records of cautiously elected terms (single words or phrases) symbolizing concepts (objects) in a particular domain. Con-

trolled vocabularies are typically structured hierarchically, and explicitly represent various conceptual relations, such as the broader- (generic), narrower- (specific) and synonymy (similar) relations. Furthermore, each concept is typically associated with a distinctive code that bestows each concept with a unique sense. The unique sense of a concept, in combination with the concept's relationship to others, makes available a clearer and more harmonized understanding about its meaning. Controlled vocabularies exist for many domains including, the procurement- (e.g., UNSPSC, ecl@ss, CPV), patent- (e.g., IPC) and medical domain (e.g., UMLS, MeSH).

As these vocabularies are available in machine-readable format, we can use them as resources in computer applications to reduce some of the ambiguity of natural language by associating pieces of information (e.g., documents) to concepts in these vocabularies. This can allow heterogeneous information to become homogenous information and can ultimately lead to intelligent organization, standardization (interoperability), and visualization of unstructured textual information. However, these resources have a clear weakness. As trained professionals typically construct and maintain these resources by hand, their content (terms denoting concepts), and representation (relations among concepts) can quickly be out-dated. Recognizing that large quantities of electronic text are available these days, it is advantageous to acquire significant terms from these collections (semi-) automatically, and to update the concepts in controlled vocabularies with this additional information. It is essential that such a technique discrimi-

¹ Also referred to as taxonomies, nomenclatures, thesauri or (light-weight) ontologies

nate well between concept-related and concept-neutral terms.

Statistical- and information-theoretic *feature selection techniques* have proved useful in the areas of information retrieval and text categorization. In information retrieval, feature selection techniques such as *document frequency* and *term frequency/inverse document frequency* (tfidf) are often adopted for sorting out relevant documents from irrelevant ones given a certain query. In text categorization, techniques like *chi-square*, *information gain*, and *odds ratio* are applied to reduce the feature set, as to allow the classifier to learn from smaller sets of relevant terms. Interestingly, despite their known ability to identify significant and discriminative terms for categories, it seems that no extensive study has been made that empirically establishes the suitability of the same techniques for the task of enhancing the content of large (semantic) knowledge sources such as controlled vocabularies.

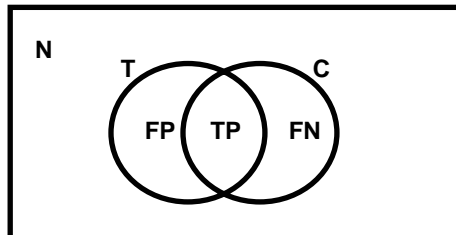
This study evaluates and compares the performance of four well-known feature selection techniques when applied to the task of detecting concept-significant terms in texts. We describe a preliminary experiment where we let each of these techniques weight and rank terms in a collection of manually labeled medical literature (OHSUMED), and we evaluate these lists of terms by comparing them against terms symbolizing 2317 concepts in the Medical Subject Headings (MeSH) vocabulary.

2 Feature Selection Techniques

In text categorization, feature selection is the task of selecting a small number terms from a set of documents that best represents the meaning of these documents. (Galavotti et al., 2000) Many techniques have been developed for this task (see Sebastiani, 2002 for an overview), and we report on four such techniques in this study. The techniques we evaluated were *chi-square*, *odds ratio* and *correlation coefficient*. We also included a metric we proposed ourselves, which we name *category frequency*.

In the following formulas, t_j denotes a term and c_i a concept, where each function assigns a score to

that term, indicating how significant that term is for that particular concept. Below, T represents documents containing t_j , and C corresponds to all documents that a professional indexer has assigned to concept c_i . TP stands for documents shared by both c_i and t_j , and FN for the set of documents belonging to c_i but not including t_j . FP represents documents that do not belong to c_i but contain t_j . N represents all documents in the text collection.



2.1 Category Frequency (cf)

We computed the category frequency as:

$$cf(t_j, c_i) = \frac{|TP|}{|C|}$$

That is, we compute the fraction of documents shared by t_j and c_i , and the total number of concept-relevant documents. We base category frequency on the notion that the significance of a term can be determined simply by establishing its distribution among the relevant documents of a concept. With this technique, we do not take the additional distributional behavior of the term into consideration. That is, this technique will not penalize terms that have a wide distribution in a text collection and it will rank terms occurring frequently among concept-relevant documents higher than terms that occur rarely in this set. We regard category frequency as baseline technique.

2.2 Odds ratio ($odds$)

The odds of some event taking place, is the probability of that event occurring divided by the probability of that event not taking place. (Freedman et al. 1991)

The rationale behind Odds ratio is that a term is distributed differently among relevant and non-relevant documents to a concept, and Odds ratio

determines whether it is equally probable that we find that term in both these sets of documents. We computed the Odds ratio according to the formula given by Mladenic (1998):

$$odds(t_j, c_i) = \frac{\frac{|TP|}{|C|} \bullet \frac{1 - \frac{|FP|}{|N-C|}}{1 - \frac{|TP|}{|C|} \bullet \frac{|FP|}{|N-C|}}}{\frac{|TP|}{|C|} \bullet \frac{1 - \frac{|FP|}{|N-C|}}{1 - \frac{|TP|}{|C|} \bullet \frac{|FP|}{|N-C|}}}$$

To be more precise, Odds ratio computes the ratio between the probability of term t_j occurring in the relevant document set of concept c_i , and the probability of t_j occurring in documents that are not relevant to c_i . Therefore, in contrast to category frequency, Odds ratio additionally considers the distribution of t_j in those documents that are not relevant to c_i , and will thereby decrease the significance of those terms that occur frequently in that set.

2.3 Chi-square (*chi*)

Chi-square measures the difference between *observed* values in some sample and values we can *expect* to observe in this sample. (Freedman et al. 1991). When we apply chi-square to perform feature selection, we assume that a term t_j and a concept c_i are independent of each other. Next, we test this hypothesis by measuring the difference between those co-occurrence relations between t_j and c_i we have observed in our text collection, and those co-occurrence relations we can expect to happen by chance. If chi-square determines that those values we have observed are significantly different from those expected values, we reject initial hypothesis and conclude that some significant relationship exists between term t_j and concept c_i .

We computed the chi-square according to the definition by given by Yang and Pedersen (1997):

$$chi(t_j, c_i) = \frac{|N| \left[\frac{|TP| \bullet |N - (T - C + TP)| - |TP| \bullet |FN|}{|T| \bullet |N - A| \bullet |C| \bullet |N - C|} \right]^2}{|N| \left[\frac{|TP| \bullet |N - (T - C + TP)| - |TP| \bullet |FN|}{|T| \bullet |N - A| \bullet |C| \bullet |N - C|} \right]^2}$$

If we detect no difference between observed and expected values, then t_j and c_i are truly independent and we obtain a value of zero for t_j . Moreover, chi-square regards terms as less significant when smaller differences are obtained, while considering

terms as more significant when bigger differences are observed.

2.4 Correlation Coefficient (*cc*)

Ng et al. (1997) offer a variant to the chi-square metric. In contrast to chi-square, correlation coefficient assigns a negative value to a term t_j when a weaker correspondence between t_j and concept c_i has been observed. Ng et al. motivate their proposed technique by saying that, if there is some suggestion that a term is significant in the relevant document set then that term is preferred over terms that are significant in both relevant and non-relevant documents. This technique diminishes the significance of terms occurring in non-relevant documents considerably, while more drastically promoting terms that frequently occur in relevant documents to a concept c_i .

$$cc(t_j, c_i) = \frac{\sqrt{|N|} \left[\frac{|TP| \bullet |N - (T - C + TP)| - |TP| \bullet |FN|}{|T| \bullet |N - A| \bullet |C| \bullet |N - C|} \right]^2}{\sqrt{|T| \bullet |N - A| \bullet |C| \bullet |N - C|}}$$

3 Experimental Setup

In this section, we explain our data and experimental methodology.

3.1 Controlled Vocabulary

Medical Subject Headings (MeSH)² is one of the more famous controlled vocabularies to date. MeSH's primary purpose is as a tool for indexing medical-related texts and it is an essential aid when searching for biomedical and other health-related literature in the Medline Database³.

MeSH is designed and updated (once-a-year) by trained professionals and it represents a large assortment of concepts from the medical domain. The latest version (2007) contains a total of 22,997 so-called *descriptors* which are terms that symbolize these concepts. Accompanying each concept is a unique identification code (so called *tree number*). This code determines the precise location of each concept in the hierarchy, and from it, we can resolve which terms give a more general definition of a particular concept (i.e., the descriptors of its ancestral concepts), which terms describe similar

² <http://www.nlm.nih.gov/mesh/>

³ <http://medline.cos.com/>

concepts (i.e., siblings concepts) and which terms denote more specific cases of a particular concept (i.e., descendant descriptors). MeSH arranges concepts in an eleven level deep hierarchical structure, defining highly generic to very specific concepts. For instance, at the second level⁴, we find 16 broad concepts, including “*Diseases*”, “*Health Care*” and “*Organisms*”. As we navigate further down the tree structure, we find increasingly more specific concepts, such as “*Respiratory Tract Diseases*” >> “*Lung Diseases*” >> “*Atelectasis*” and >> “*Middle Lobe Syndrome*”. Additionally, many of the concepts in MeSH have *entry terms* associated with them. These are additional terms being synonyms (or quasi-synonyms, such as different spellings and plural forms) to the descriptor. E.g., we find that *cancer*, *tumor*, *neoplasms* and *benign neoplasm* are all entry terms for the concept “*Neoplasm*”.

The OHSUMED collection, that we describe in the next section, included relevance judgments for 4904 MeSH concepts. We included 2317 of these concepts in our experiment, each with a unique location in MeSH, and their descriptors became our gold standard. We considered each descriptor (e.g., “*Lung Diseases*”) of a concept as a significant term for that concept, composed with the descriptors of its descendants (e.g., “*Atelectasis*” and “*Middle Lobe Syndrome*”). If a concept was a leaf (e.g., “*Middle Lobe Syndrome*”), instead we additionally regarded each (possible) entry term (e.g., *brock syndrome*, *brocks syndrome*, *brock's syndrome*) as significant for that particular concept.

3.2 Text Collection

The textual resource used in these experiments was the OHSUMED collection (Hersh, 1994). OHSUMED is a subset of the Medline Database and includes 348.566 references to 270 medical journals collected between 1987 and 1991. Most of these texts are references to journal articles, but some are references to conference proceedings, letters to editors and other medical reports. While many references include only a title, the majority also include an abstract, truncated at 250 words. We set the content of a document to include the title and (possibly) the abstract of a reference. In

⁴ We added a root node in these experiments to connect all branches.

view of the fact that OHSUMED includes references from Medline, each reference consequently came with a number of manually assigned MeSH concepts. That is, for each of the 2317 concepts previously selected, we knew their relevant and non-relevant document sets.

Before indexing this collection, we performed inflectional stemming and NP chunking, and we omitted all terms not identified as single nouns or noun phrases. Once the indexing was complete, we applied each feature selection technique to the 2317 features sets. We setup this process as follows: Given a MeSH concept, we retrieved all of its associated documents from the document collection, and collected the complete feature set of (unique) terms. In order to contrast these terms with those terms we had in our gold standard, we kept only the ones that were already present (or parts of descriptors) in MeSH. While these lists typically included 1400 terms, for some concepts we obtained over 5000 terms, while for others we obtained less than 100. Next, we applied each feature selection technique to weight and rank each of these terms. Once this process was complete, we obtained four lists for each of the 2317 concepts, where each list included the same set of terms, while varying only in respect to the ordering of those terms. Next, we evaluated each feature selection technique by comparing the lists they had produced with terms in our gold standard we knew where significant.

4 Evaluation Metrics

We evaluated the performance of each feature selection technique based on the ordered feature lists previously obtained. Essentially, a technique was performing well if it ranked significant terms higher than less significant terms. We employed three evaluation metrics: the *Wilcoxon rank-sum test*, *precision at n*, and the *Spearman rank correlation*.

4.1 Wilcoxon Rank-Sum Test

Using the Wilcoxon Rank-Sum test⁵ (Mann and Whitney, 1947), we measured the overall tendency of each technique ranking significant terms either

⁵ Alternatively, Mann-Whitney U test.

high or low. This metric took an ordered list of terms for a given concept, and verified whether significant terms normally appeared at the beginning or at the end of this list. The rank sum becomes low when significant terms exist near the beginning of the list and high when insignificant terms precede relevant terms in the list. We considered the ordering of terms as non-random when the sum of the ranks varied more than we could expect by chance.

4.2 Precision at n

Precision at n also provides a mean for measuring the quality of rankings. In contrast to the previous metric, we can inspect the precision at certain positions in this ranking. Precision at n gives the accuracy obtained for the first n terms that we know from our gold standard to be significant. A perfect technique therefore places all significant terms at the beginning of the list, while positioning less significant terms at the lower end of the list. We computed precision at n ($p(n)$) according to:

$$p(n) = \frac{rel_n}{n}$$

where n is some ranking position and rel_n the number of relevant terms found among the first n terms suggested. We computed the precision at rank positions 5, 10, 15, 20, 30, 100, 200, 500 and 1000, and by averaging the precision values for each technique over all 2317 concepts.

4.3 Spearman's Rank Correlation

Semantic similarity⁶ measures are metrics for computing the relatedness in meaning between concepts (or terms denoting them) based on their distance to each other in a hierarchy. (Budanitsky and Hirst, 2004). They all build upon the assumption that concepts (or terms denoting them) situated closely in the hierarchical space are more similar in meaning than concepts (or terms denoting them) that are separated farther away. E.g., in WordNet (Fellbaum, 1998), we find that *wolf* and *dog* are more related than *dog* and *hat*, since, in WordNet, *wolf* and *dog* share the same parent (i.e., *Canine*).

⁶ Also known as semantic distance or relatedness.

The idea was to compare the ordering of terms decided by each feature selection technique, with the order these terms obtained based on their *semantic distance* to respective concepts in our experiment. That is, let's suppose that some technique determined '*hypoglycemia*' to be insignificant for the concept "*Diabetes Mellitus*", and thereby giving it a low rank. However, if we compute the distance between '*hypoglycemia*' and "*Diabetes Mellitus*", in MeSH, we find that '*hypoglycemia*' gets a high relatedness value, as this term symbolizes one of two siblings of "*Diabetes Mellitus*" and thereby receives a high rank. If cases like this were frequent, it would indicate that this particular technique was unable to detect significant terms.

Spearman's Rank Correlation (*rho*) is a metric for comparing ordering of items. When two lists come in the same order, they are identical, and the rank correlation becomes one (1). Conversely, if one is the inverse of the other, then the correlation becomes -1. We obtain a correlation value of zero when there is no relation between the two. The rank correlation is computed using:

$$rho = \frac{6 \sum d_i^2}{n \bullet (n^2 - 1)}$$

where d_i is the difference between each entry pair, and where n equals the number of entry pairs.

Using Leacock-Chodorow's measure of path length (Leacock and Chodorow, 1994), we computed the distance between each term in our feature lists and a concept in question. We now had two orderings with the identical set of terms, which we could compare. Specifically, one list including the ordering of terms decided by some feature selection technique, and the other being a list based on the semantic distance between each term and a certain concept.

In hierarchies such as MeSH, relatedness rapidly decreases as distance increases. This is especially true when a path between a term and a concept leads through the root of the hierarchy. These are cases when a term and a concept are positioned in separate branches of the 16 main concepts at the second level. Recognizing this fact, we (additionally) normalized the path length metric by setting a

threshold, such that the relatedness value of became zero if the path from a term to a concept included the root concept.

5 Results

The Wilcoxon Rank Sum test gave us a clear indication that, for a large majority of concepts, each of the four feature selection techniques ranked significant terms before less significant terms. Further, Table 1 illustrates the precision that each feature selection technique obtained at each of the nine ranking positions, where these values are averaged over all 2317 concepts. We observe that Odds ratio (*odds*) scores the lowest precision values at all cut-off points on this test. Both the Chi-square (*chi*) and Correlation Coefficient (*cc*) metrics perform better than the rivaling techniques. In fact, their performances are identical. Our baseline technique (*cf*) performs slightly lower than *chi* and *cc*.

Rank position	Feature Selection Technique			
	<i>cf</i>	<i>odds</i>	<i>chi</i>	<i>cc</i>
5	0,32	0,23	0,39	0,39
10	0,19	0,17	0,25	0,25
15	0,14	0,13	0,19	0,19
20	0,12	0,11	0,16	0,16
30	0,09	0,08	0,12	0,12
100	0,04	0,04	0,05	0,05
200	0,02	0,02	0,03	0,03
500	0,01	0,01	0,01	0,01
1000	0,009	0,009	0,009	0,009

Table 1: Precision at rank position 5 --1000. Values are averaged over 2317 experiments.

In Table 2, we see the average correlation in rankings between the lists of terms ordered by each technique, and the ordering of the same set of terms based on their semantic distance to respective concepts included in these experiments. Here, we assigned the real distance value of a term even if its path to a concept included the root concept. Again, values are averaged over all 2317 concepts.

Feature Selection Technique	Rank Correlation
<i>cf</i>	0,30
<i>odds</i>	-0,19
<i>chi</i>	-0,06
<i>cc</i>	-0,13

Table 2: Rank Correlations averaged over 2317 concepts. Path via root node allowed.

This tells us that, *chi*, *cc* and *odds* all have a tendency toward ranking terms in *contradictory* order to the Leacock-Chodorow's measure of semantic distance. Contrastively, we observe a positive correlation between our baseline technique (*cf*) and that distance measure, although this correlation is on the weaker end of the scale. This indicates that *cf* more often ranked closely positioned terms to our concepts higher, than it ranked terms situated more distantly from our concepts in MeSH.

When we normalized the Leacock-Chodorow measure, we obtained positive correlation value for all techniques and they came to conform more to each other. (Table 3)

Feature Selection Technique	Rank Correlation
<i>cf</i>	0,35
<i>odds</i>	0,20
<i>chi</i>	0,19
<i>cc</i>	0,16

Table 3: Rank Correlations averaged over 2317 concepts. Paths via root node given a value of zero.

6 Discussion

We have evaluated and compared four feature selection techniques on the task of detecting significant terms for concepts in the medical domain. Our results suggest that all techniques behave similarly in respect to ranking significant terms. Both the Wilcoxon rank-sum test and precision at *n* gave a clear indication of this. Although we evaluated each feature selection technique on nine different ranking positions, it probably makes more sense to do it only on ranking positions 5—20. We can imagine a controlled vocabulary editor getting a

list of suggested terms to add to the terminology. In such a scenario, it is likely that the editor is only interested in verifying the relevance of 2—15 terms. Failing to notice significant terms appearing later in the list should be a minor concern.

However, we observed noticeable differences between the techniques when we compared their ordered set of terms with the semantic relatedness values of these terms. Results showed that the simplest technique (*cf*) conform to the conceptual relations among terms in MeSH the most, while the more sophisticated techniques tended to rank terms in contradictory order. We are aware that these results can be different if we choose some other semantic similarity metric. However, to the best of our knowledge, evaluating feature selection techniques using semantic similarity measures has never been tested and we consider semantic relatedness measures as interesting alternatives to the other evaluation metrics and they should provide us with some additional information regarding the behavior of feature selection techniques. In the future, we intend to investigate the justifications of semantic similarity measures and the role these measures can have in our setting.

What our study boils down to is that of determining whether the task we appoint to feature selection techniques in this setting is different from, similar or even identical to the task these techniques are intended to solve in text categorization. At this point, we cannot provide a straightforward answer to that question. It is reasonable to argue that the tasks are similar if we employ these techniques in some (semi-) automated scenario, where it is an absolute necessity that top ranking terms have high discriminating power. However, if these techniques are only part of, say, some editing tool where trained professionals can judge the outcomes, then we might want to consider the tasks as different.

References

Alexander Budanitsky and Graeme Hirst. 2004. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics* 32(1):13—47.

Christiane D. Fellbaum. 1998. *WordNet, an electronic lexical database*. MIT Press.

David Freedman, Robert Pisani, Roger Purves, and Ani Adhikari. 1991. *Statistics*. Second edition. Norton. New York.

Luigi Galavotti, Fabrizio Sebastiani, and Maria Simi. 2000. Experiments on the use of feature selection and negative evidence in automated text categorization. *Proceedings of ECDL-00, 4th European Conference Research and Advanced Technology for Digital Libraries*.

William Hersh. 1994. Ohsumed: An interactive retrieval evaluation and new large test collection for research. *Proceedings of the 17th Annual Intl. ACM SIGIR Conference on R&D in Information Retrieval*.

Claudia Leacock and Martin Chodorow. 1998. Combining Local Context and WordNet Similarity for Word Sense Identification. *WordNet: An Electronic Lexical Database*. C. Fellbaum, MIT Press: 265—283

Dunja Mladenic. 1998. Feature Subset Selection in Text-Learning. *European Conference on Machine Learning*.

Hwee T. Ng, Wei B. Goh and Kok L. Low. 1997. Feature selection, perceptron learning, and a usability case study for text categorization. *Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval*.

Fabrizio Sebastiani. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.* 34(1): 1—47.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. *Proceedings of ICML-97, 14th International Conference on Machine Learning*.