Learning-based Detection of Scientific Terms in Patient Information

Veronique Hoste, Els Lefever, Klaar Vanopstal, Isabelle Delaere

LT3 Language and Translation Technology Team
University College Ghent
Groot-Brittanniëlaan 45
9000 Gent, Belgium
veronique.hoste, els.lefever, klaar.vanopstal, isabelle.delaere@hogent.be

Abstract

In this paper, we investigate the use of a machine-learning based approach to the specific problem of scientific term detection in patient information. Lacking lexical databases which differentiate between the scientific and popular nature of medical terms, we used local context, morphosyntactic, morphological and statistical information to design a learner which accurately detects scientific medical terms. This study is the first step towards the automatic replacement of a scientific term by its popular counterpart, which should have a beneficial effect on readability. We show an F-score of 84% for the prediction of scientific terms in an English and Dutch EPAR corpus. Since recasting the term extraction problem as a classification problem leads to a large skewedness of the resulting data set, we rebalanced the data set through the application of some simple TF-IDF-based and Log-likelihood-based filters. We show that filtering indeed has a beneficial effect on the learner's performance. However, the results of the filtering approach combined with the learning-based approach remain below those of the learning-based approach.

1. Introduction

Despite the efforts of the regulatory authorities to produce guidelines which stipulate that "all technical terms should be translated into a language which is understandable for patients", patients are still confronted with incomprehensible information. Previous research (Van Vaerenbergh, 2007) has shown that the use of scientific terminology is one of the factors which greatly influences the readability of this patient information. The leaflet for the public is mostly an adaptation of the scientific leaflet due to the legal requirement that the leaflet is closely related to the so-called product summary meant for experts, and therefore also written in expert language, with expert terminology.

In this paper, we address the problem of scientific term detection in a patient information corpus. Automatic term extraction is crucial in many domains of (computational) linguistics, including automatic translation, text indexing, the automatic construction and enhancement of lexical knowledge bases, etc. In research on automatic term extraction, two different directions mainly have been taken. On the one hand, the linguistic-based or rule-based approaches, e.g. (Dagan and Church, 1994), (Ananiadou, 1994), (Fukuda et al., 1998) make use of hand-coded rules and look for specific (mostly language-specific) linguistic structures that match a number of predefined syntactic patterns. On the other hand, the statistical corpus-based approaches, e.g (Pantel and Lin, 2001), (Andrade and Valencia, 1998), extract terms using different types of metrics to measure the information between words. Along the same corpus-based line, different machine learning approaches have been proposed using learning techniques such as Hidden Markov Models (Collier et al., 2000) or Support Vector Machines (Kazama et al., 2002), and combination methods such as boosting (Vivaldi et al., 2001), etc. on feature sets encoding lexical, POS, orthographic, and other possibly relevant information. Hybrid approaches combining both linguistic and statistical information have also emerged, e.g. (Maynard and Ananiadou, 1999), (Frantzi and Ananiadou, 1999). For an overview of the field, we refer to (Hirshman et al., 2002) and (Ananiadou and McNaught, 2006).

Although most term extraction research in the biomedical domain is focused on recognizing gene and protein names, etc., the techniques developed in this domain are also useful for the specific problem of medical term detection in patient-oriented information. Since there are no (or very limited) lexical resources available which distinguish between the scientific and popular use of medical terms, we could not rely on these lexicons for our task at hand. This implies, for example, that the MeSH heading *headache* [C10.597.617.470], makes no distinction between the ¡synonymous scientific terms *cephalalgia*, *cephalgia*, *hemicrania* and the more popular term *head pain*. In order to bypass the problem of lacking useful dictionaries, we will use a machine learning approach on a variety of information sources. In earlier work (Hoste et al., 2007), we contrasted a lexicon-based with a simple learning-based approach which did not rely on any external lexical resources and showed that a learning-based approach outperforms the lower recall lexicon-based approach. In this paper, we further experiment with different types of features and investigate some filtering techniques to reduce the skewedness of our data sets.

The remainder of this article is structured as follows. Section 2. presents the EPAR corpora, describes the corpus annotation and introduces the learning method which will be used in all experiments. Section 3. gives an overview of the different types of information sources which will be incorporated in the feature vectors. Section 4. describes the results of the experiments on the English and Dutch EPAR data sets, whereas Section 5. focuses on the problem of the data set skewedness. Section 6. concludes and points to fu-

ture work.

2. Experimental setup

In order to quantify and automatically detect the use of scientific terminology in Dutch and English medicinal texts, we collected two data sets of EPAR summaries from the EMEA (European Medicines Agency), one for each language. EPAR stands for "European Public Assessment Report" and is a text which is prepared at the end of every centralized evaluation process. Although these EPAR abstracts were originally intended to provide information comprehensible to the general public, they suffer from the same shortcomings as the package leaflets which are also often considered too technical.

2.1. Corpus annotation

For both Dutch and English, we collected a parallel corpus of 317 EPAR summaries. This corpus was used to calculate the TF-IDF and Log-likelihood statistics as described in the following section. 20 summaries of each language were manually annotated (English: 17,502 tokens; Dutch: 17,098 tokens) by two linguists, who annotated the corpora in parallel. As input, they received free text, which was tokenized and provided with lemmatization and part-of-speech information. Tokenization was performed by a rule-based system using regular expressions. Part-of-speech tagging and text chunking for English was performed by the memory-based tagger MBT (Daelemans et al., 2003), which was trained on text from the Wall Street Journal corpus in the Penn Treebank (Mitchell et al., 1993), the Brown corpus (Kucera and Francis, 1967) and the Air Travel Information System (ATIS) corpus (Hemphill et al., 1990). Part-of-speech tagging for Dutch was again performed by the memory-based tagger MBT, this time trained on the Spoken Dutch Corpus (CGN)[1]

The annotators had to differentiate between the following three coarse-grained labels: (i) 'scientific' for real scientific terms, (ii) 'medium' for terms that are used with a specific medical meaning or consecutive terms that form together frequently used medical expressions and (iii) 'popular' for all general vocabulary terms. Overall, both annotators gave a scientific tag to about 10% of all tokens. The kappa scores are 0.64 (English) and 0.76 (Dutch).

2.2. Learning-based Scientific Term Extraction

Earlier experiments (Hoste et al., 2007) with lexicon-based term extraction revealed high precision scores, as opposed to recall scores below 50% for both languages. The existing lexicons suffer from two main shortcomings: (i) their coverage remains low, especially for Dutch and (ii) they do not make a distinction between popular and scientific medical terms, which is our main goal. In order to overcome these shortcomings, we integrated local context, morphological, morpho-syntactic and lexical information in a machine learning approach to scientific term extraction.

[1]More information on this corpus can be found at http://lands.let.ru.nl/cgn/.

We used the TIMBL (Daelemans and van den Bosch, 2005) software package that implements a version of the k nearest neighbour algorithm. It is an implementation of the IB1 (Aha et al., 1991) algorithm, with some additional features (such as different metrics for the calculation of the distances between two items). An MBL system consists of two components: a memory-based learning component and a similarity-based performance component. During learning, the learning component adds new training instances to the memory without any abstraction or restructuring. During classification, the classification of the most similar instance in memory is taken as classification for the new test instance. In other words, given a set of instances or data points in memory: (x_1, y_1) (x_2, y_2) (x_3, y_3) ... (x_n, y_n), the task at classification time is to find the closest x_i for a new data point x_q. In order to do so, the following components are crucial: (i) a distance metric which looks at the number of matching and mismatching feature values in two instances, (ii) the number of nearest neighbours to look at and (iii) a strategy of how to extrapolate from the nearest neighbours.
The learner had to differentiate between two classes: "scientific" and "popular". The scientific category is our category of interest and represents the terms which one of both annotators labeled as scientific. All other words in the text were considered "popular". We performed 20-fold cross-validation on the data sets, 20 being the total number of annotated documents per language.

3. Information sources

The following information sources were incorporated in the feature vector.

3.1. Local context information

We included word form, lemma and part-of-speech information of two words to the left and two words to the right of the focus word. This local context information can contain specific phrases which signal the presence of a term, for example "is referred to as", "denotes", "is defined as", "is called", "known as" etc. (see for example (Pearson, 1996)). Another typical indication in the local context of the presence of a term is the use of brackets. This can be a full form with its corresponding abbreviation between brackets:

> (...) **body mass index (BMI)** greater or equal to 30 kg/m
> (...) treatment of **protease inhibitor (PI)** experienced HIV-1 infected adults and children above the age of 4 years.
> (...) a specific type of receptors, the **cannabinoid type 1 (CB1)** receptors

Furthermore, brackets may also indicate explanations of scientific terms:

> The receptors are also found in **adipocytes (fat tissue)**.
> (...) compared the effect of ACOMPLIA with that of a **placebo (dummy treatment)** on weight loss over one to two years.
> Actraphane may cause **hypoglycaemia (low**

blood glucose).

3.2. External lexicons

As lexical features, we relied both on external lexicons and corpus-specific lexical features. For English, we started from the MeSH lexicon and the English sections of Taalvlinder and Ziekenhuis.nl (Hoste et al., 2007). In addition to these sources, we used the Specialist Lexicon from UMLS. This lexicon covers both the English general language and concepts from the field of biomedicine and was originally designed to support the SPECIALIST Natural Language Processing System and to generate indexes to the Metathesaurus. The combined English sources resulted in a lexicon containing 588,756 general and scientific terms. In order to filter out general vocabulary terms we isolated the terms which were also present in Celex lexical database. This reduced the number of English terms to 559,404 unique scientific terms.
As mentioned above, the Dutch lexicons (Taalvlinder and Ziekenhuis.nl) used in previous experiments had a rather low coverage. Therefore, we incorporated several additional Dutch medical sources:

- **ICD-9 DE**: the Dutch translation of ICD-9 CM. (Ninth Revision of the international Classification of Diseases, Clinical Modification). This classification is used to code and classify mortality data from death certificates and is derived from the World Health Organization's ICD-9 classification. It comprises 7249 terms.
- **Elseviers Medische Encyclopedie**: a Dutch medical encyclopedia intended for the general audience, containing 6004 scientific and popular index terms.
- **Gezondheid.nl**: an online medical encyclopedia containing 5691 lemmas.
- **Wikipedia**: the medical entries listed in "Gezondheid van A tot Z", the alphabetical index of subjects related to health care in Wikipedia. This source provided us with a total of 674 unique terms.

A total of 26,324 unique Dutch terms were obtained from this collection of lexicons. The intersection with Celex resulted in 22,606 unique scientific terms.

3.3. Corpus-specific lexicon features

Quite some research has been done in order to detect words that are specific to a corpus based on corpus comparison. Consequently, a wide range of different techniques have been developed in information retrieval as well as in the field of computational terminology (e.g. (Salton, 1989), (Dunning, 1993)). (Salton, 1989) has tried to determine the weight of a word (in a collection of documents) by calculating TF-IDF scores, whereas other researchers such as (Dunning, 1993) and (Rayson and Garside, 2000), have explored the use of the Log-likelihood measure to discover keywords which differentiate between corpora. Next to that, techniques of Mutual Information (Church and Hanks, 1990) and hypergeometric distribution ((Lafon, 1980), (Lebart and Salem, 1994)) have been explored for finding lexicon-specific terms. We considered both TF-IDF and Log-likelihood to expand our feature set with corpus-specific lexicon features.

The TF-IDF (term frequency inverse document frequency) statistic (Salton, 1989) combines two hypotheses: a search term is of more value when it occurs in few documents (IDF) and distinctive terms have a high frequency in a given document (TF). As we also need to pin-point distinctive keywords (scientific terms in our case), we calculated TF-IDF for all terms in the full EPAR corpus. For calculating the IDF, we enlarged the English EPAR corpus with all written and spoken documents of the British National Corpus (BNC). For Dutch, the Twente News Corpus (TNC)[2] was taken as reference corpus. Calculating TF-IDF on the EPAR terms should enable us to extract lexicon specific scientific terms that have much lower frequencies in a balanced reference corpus such as the BNC and TNC.
Given a document collection D, a word w, and an individual document $d \in D$,

$$W_d = f_{w,d} * log(|D|/f_{w,D}) \quad (1)$$

where $f_{w,d}$ equals the number of times w appears in d, $|D|$ is the size of the corpus and $f_{w,D}$ equals the number of documents in which w appears in D (Berger et al., 2000). In order to determine the TF-IDF threshold for considering a term as being scientific, we performed 20-fold cross-validation on the labeled EPAR corpus. These results are shown in Figure 1.

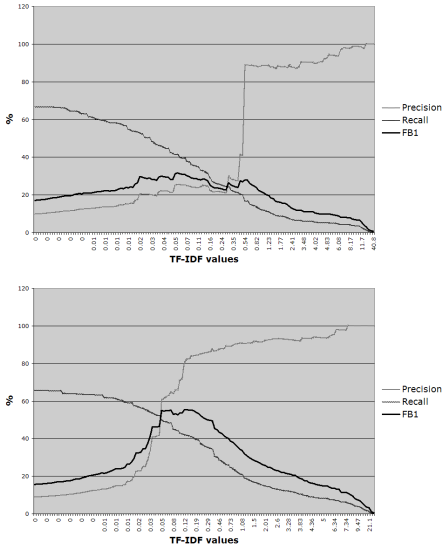

Figure 1: Cross-validated TF-IDF values for English (above) and Dutch (below)

This led to the selection of 0.1, 0.2, and 0.5 as thresholds for both languages, values which were then used to create a binary TF-IDF feature. We also experimented with these threshold values to rebalance the highly skewed data set

[2]Available at: http://wwwhome.cs.utwente.nl/∼druid /TwNC/TwNC-main.html

(as we will explain in Section 5.).

As a second measure, we calculated **Log-likelihood**. Both (Daille, 1995) and (Kilgarriff, 2001) have determined empirically that LL is an accurate measure to find the most "surprisingly" frequent words in a corpus that also corresponds fairly well to what humans might associate with distinctiveness of terms. We first produced a frequency list for each corpus and calculated the Log-likelihood statistic for each word in this frequency list. In the formula below, N corresponds to the number of words in the corpus, whereas the "observed values" O correspond to the real frequencies of a word in the corpus. The formula to calculate both the expected values (E) and the Log-likelihood have been described in detail by (Rayson and Garside, 2000).

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \quad (2)$$

We used the resulting Expected values to calculate the Log-likelihood:

$$-2ln\lambda = 2\sum_i O_i ln(\frac{O_i}{E_i}) \quad (3)$$

Manual inspection of the Log-likelihood figures confirmed our hypothesis that scientific terms in our EPARs usually get assigned high LL-values (combined with low BNC or TNC frequencies). The Log-likelihood information was integrated as a binary feature. Terms with Log-likelihood value above a predefined threshold and with BNC/TNC frequency below a predefined threshold were set to 1, the others were set to 0. Both thresholds were validated on the EPAR corpora using 20-fold cross-validation. For both Dutch and English, the thresholds were set to 2 (LL value) and 1000 (BNC/TNC frequency).

3.4. Affixes, orthographic features and trigrams

Affixation and (semi-)neoclassical compounding have proved to be extremely productive word formation techniques since the 16th century. Greek and -especially- Latin were the languages of science, leaving very distinct traces in present-day terminology. The use of these Greco-Latinates has some advantages over the use of vernacular terms: they create terminological continuity and consequently increase the efficiency of medical communication. However, the overall comprehensibility of these Greco-Latinate forms to the general audience is low.

Therefore, we incorporated Latin and Greek affixes as one of the criteria to detect scientific medical terms. A list of prefixes, suffixes and confixes compiled by (Banay, 1948) was completed during an experimental analysis of MeSH terms (Vanopstal and Van Wiele, 2007). For English, this list contains 745 prefixes and 17,520 terms. The corresponding figures for Dutch are 683 and 8713. In this list, confixes which occur in initial position are considered as prefixes and confixes in final position as suffixes. From this list of affixes, three additional features were deduced: the presence of a prefix, the presence of a suffix and the presence of both a prefix and a suffix in one term.

Orthographic features are to inform us about the presence or absence of numeric symbols and of the use of multiple capital letters, which indicates the use of abbreviations or acronyms. Furthermore, we included two trigram features which represent the initial and final trigram of a given word.

4. Experimental results

Table 1 gives an overview of the 20-fold cross-validation results of TIMBL on the English and Dutch EPAR data sets. The accuracy results are measured on the complete data set. The high accuracy scores (>90%) can partially be explained by the highly skewed class distribution in the data set. If the number of negative and positive instances is highly unbalanced, this will typically lead to a classifier which has a low error rate for the majority class and a high error rate for the minority class. Since about 90% of the words in the EPAR corpus are non-scientific terms, high precision scores can be obtained even without detecting any scientific term. The last three columns list the precision, recall and FB1 results on the scientific terms, our category of interest. Overall, we can observe an F-score of 84% for the detection of scientific terms in the English EPARs. The corresponding result for the Dutch EPARs is 85%. Furthermore, we can observe for both languages that the precision scores are consistently higher than the recall scores.

If we consider the contribution of the different types of features, the following can be observed. The local context features, with inclusion of lemma, word form and part-of-speech information of the focus word give the best results. If we leave out the information on the focus word, there is an expected drop in performance of about 20%, leading to an F-score of 56% for both languages. The three lexical features (external lexicon, TF-IDF and Log-likelihood) show high precision scores for English, but suffer from a low coverage, i.e. 95% precision versus 24% recall. These results are in line with earlier results on the low coverage of lexicon-based approaches (see for example (Aubin and Hamon, 2006)). For Dutch, however, these scores are much more balanced: 65% precision versus 63% recall, as also shown earlier in Figure 1. Finally, we can observe a highly beneficial effect on precision of the morphological information on prefixes, suffixes, trigrams, capitalization and word-internal numbers.

5. Rebalancing the data sets

Taking into account the highly skewed data set (about 10% of scientific terms) which might cause the learner to have a bias towards the majority non-scientific class, we experimented with several filters on the data. These filters should not only result in a more balanced data set, but also lead to a reduction of the training data. In order to evaluate the effect of this filtering, the test set is kept stable. The filters are applied to the test set as follows: one part is automatically classified by the filter, whereas the remaining part is handled by the classifier which is trained on the reduced, more balanced data set.

We experimented with three TF-IDF filters and one LL filter, both also being incorporated in the feature vectors described in Section 3. The results in Table 2 show that filtering the training data leads to a large reduction of the train-

ENGLISH	Accuracy	Precision	Recall	FB1
Complete system				
All features	96.18	87.18	81.40	84.19
Baseline systems using groups of features				
Local context	94.35	81.85	70.48	75.74
Local context without focus	89.57	59.02	54.11	56.46
Lexical	90.33	94.93	23.95	38.25
Morphological	93.51	86.78	56.72	68.60

DUTCH	Accuracy	Precision	Recall	FB1
Complete system				
All features	97.25	89.38	86.93	88.14
Baseline systems using groups of features				
Local context	94.57	82.47	68.46	74.81
Local context without focus	90.07	58.61	53.25	55.80
Lexical	91.62	64.83	63.09	63.95
Morphological	94.32	89.53	58.62	70.85

Table 1: 20-fold cross-validation results on the English and Dutch EPAR data set with the complete feature vector. Contribution of the different types of feature information.

ing material and a better classifier, which is more tailored to the "scientific" class. This effect is most prominent for English, where we can observe a 7% performance increase (84% without filtering versus 91.49% with the TF-IDF $<$ 0.1 filter). Despite the fact that filtering for Dutch is more accurate at detecting "scientific" instances, the performance increase for Dutch is lower (88% without filtering versus 91.8% with the TF-IDF $<$ 0.5 filter). However, since the other part of the instances is handled by the filter, we can observe for both languages that filtering has a negative effect on classification results. As an alternative to this harsh filtering approach, we will investigate in future work how cost-sensitive classifiers, e.g. (Domingos, 1999), which set a high cost to the misclassifications of a minority class can be used for our data. Furthermore, since we are only interested in the "scientific" class, we will also investigate if we can consider our task as a one-class classification task (see for example (Manevitz and Yousef, 2001)).

6. Conclusion

In this paper, we investigated the use of a machine learning approach to scientific term detection in patient information. We showed an F-score of above 84% for the prediction of scientific terms in an English EPAR corpus. For Dutch, an 88% F-score was obtained. In future experiments, we plan to further enrich the feature vectors with cognate information and by exploiting the parallel EPAR corpora. In a second experiment, we investigated whether filtering could have a beneficial effect on the classification results on the highly skewed data sets. Our results indeed confirm that classification results benefit from a more balanced data set. We plan to further experiment with cost-sensitive classifiers and cost-sensitive classification.

In future experiments, we also plan to automatically replace the detected scientific terms by their popular counterparts. By means of a readability test, we will investigate whether this indeed leads to an improved readability of patient information.

7. References

D.W. Aha, D. Kibler, and M. Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6:37–66.

S. Ananiadou and J. McNaught. 2006. *Text mining for biology and biomedicine*. Artech House, Inc.

S. Ananiadou. 1994. A methodology for automatic term recognition. In *Proceedings of the 15th conference on computational linguistics*, pages 1034–1038.

M. Andrade and A. Valencia. 1998. Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *BioInformatics*, 4(7).

S. Aubin and T. Hamon. 2006. Improving term extraction with terminological resources. In *Advances in Natural Language Processing 5th International Conference on NLP, FinTAL 2006*.

G.L. Banay. 1948. An introduction to medical terminology i. greek and latin derivations. *Bulletin of the Medical Library Association*, 1(36):1–27.

A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. 2000. Bridging the lexical chasm: Statistical approaches to answer finding. In *Proc. Int. Conf. Research and Development in Information Retrieval*, pages 192–199.

K. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

N. Collier, C. Nobata, and J. Tsujii. 2000. Extracting the names of genes and gene products with a hidden markov model. In *Proceedings of COLING-2000*, pages 201–207.

W. Daelemans and A. van den Bosch. 2005. *Memory-based Language Processing*. Cambridge University Press.

W. Daelemans, J. Zavrel, A. van den Bosch, and K. van der Sloot. 2003. Memory based tagger, version 2.0, reference guide. Technical Report ILK Technical Report - ILK 03-13, Tilburg University.

I. Dagan and K. Church. 1994. Termight: identifying and

	Data set size			Results				
ENGLISH	**Initial**	**% inst.**	**% scient**		**Accuracy**	**Precision**	**Recall**	**FB1**
No filter	17,502	100	12.50		96.18	87.18	81.40	84.19
TF-IDF filter (<0.1)	5,995	34.25	25.20	TIMBL	95.83	94.18	88.95	91.49
				filter+TIMBL	94.70	94.18	61.43	74.36
TF-IDF filter (<0.2)	5,721	32.69	22.95	TIMBL	96.19	94.19	88.88	91.46
				filter+TIMBL	93.75	94.19	53.34	68.11
TF-IDF filter (<0.5)	3,070	17.54	33.91	TIMBL	94.00	93.66	89.20	91.38
				filter+TIMBL	92.69	93.66	44.56	60.39
LL filter	4,161	23.77	26.12	TIMBL	93.22	89.42	83.99	86.62
				filter+TIMBL	92.10	89.42	41.73	56.90

	Data set size			Results				
DUTCH	**Initial**	**% inst.**	**% scient**		**Accuracy**	**Precision**	**Recall**	**FB1**
No filter	17,098	100	11.77		97.25	89.38	86.93	88.14
TF-IDF filter (<0.1)	3,482	20.36	44.77	TIMBL	89.83	90.46	86.40	88.39
				filter+TIMBL	95.27	90.46	66.92	76.93
TF-IDF filter (<0.2)	3,200	18.72	46.34	TIMBL	91.03	90.57	90.02	90.29
				filter+TIMBL	95.22	90.57	66.32	76.57
TF-IDF filter (<0.5)	2,909	17.01	43.42	TIMBL	92.92	92.04	91.61	91.83
				filter+TIMBL	94.41	92.04	57.48	70.76
LL filter	4,399	25.73	41.99	TIMBL	87.34	90.36	78.18	83.83
				filter+TIMBL	95.77	90.36	71.73	79.98

Table 2: 20-fold cross-validation results on the English and Dutch EPAR data set. The training data sets are reduced through filtering (columns 1 to 3); the test data sets remain unchanged. The results reported in the last four columns are (i) the TIMBL classification results on the test data which are not automatically handled by the filter and (ii) the combined filtering and classification results on the complete test set.

translating technical terminology. In *Proceedings of Applied Language Processing*, pages 34–40.

B. Daille. 1995. Combined approach for terminology extraction: lexical statistics and linguistic filtering. Technical Report 5, Lancaster University: UCREL.

P. Domingos. 1999. Metacost: A general method for making classifiers cost sensitive. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pages 155–164.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

K.T. Frantzi and S. Ananiadou. 1999. The c-value/nc-value domain independent method for multiword term extraction. *Journal of Natural Language Processing*, 6(3):145–180.

K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. 1998. Toward information extraction: Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 707–718.

C.T. Hemphill, JJ. Godfrey, and G.R. Doddington. 1990. The atis spoken language system pilot corpus. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 96–101.

L. Hirshman, J.C. Park, J. Tsujii, L. Wong, and C.H. Wu. 2002. Accomplishments and challenges in literature data mining for biology. *BioInformatics Review*, 18(12):1553–1561.

V. Hoste, K. Vanopstal, and E. Lefever. 2007. The automatic detection of scientific terms in patient information. In *Proceedings of A Workshop on Acquisition and Management of Multilingual Lexicons. In conjunction with RANLP-2007.*

J. Kazama, T. Makino, Y. Ohta, and J. Tsujii. 2002. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL Workshop on NLP in the Biomedical Domain*, pages 1–8.

A. Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.

H. Kucera and W.N. Francis. 1967. *Computational analysis of present-day English*. Brown University Press, RI.

P. Lafon. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *MOTS*, 1:128–165.

L. Lebart and A. Salem. 1994. *Statistique textuelle*. Dunod.

L.M. Manevitz and M. Yousef. 2001. One-class svms for document classification. *Journal of Machine Learning Research*, 2:139–154.

A. Maynard and S. Ananiadou. 1999. Identifying contextual information for multi-word term extraction. In *Proceedings of Terminology and Knowledge Engineering Conference-99*, pages 212–221.

P. M. Mitchell, B Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

P. Pantel and D. Lin. 2001. A statistical corpus-based term extractor. In *Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, pages 36–46.

J Pearson. 1996. Strategies for identifying terms in spe-

cialised texts. Technical Report 16, Irish Association for Applied Linguistics.

P. Rayson and R. Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora, 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, pages 1–6.

G. Salton. 1989. *Automatic text processing: the transformation, analysis and retrieval of information by computer*. Addison Wesley.

L. Van Vaerenbergh. 2007. Wissensvermittlung und anweisungen im beipackzettel. zu verstandlichkeit und textqualitat in der experten-nichtexperten-kommunikation. In *Kommunikation in Bewegung. Multimedialer und multilingualer Wissenstransfer in der Experten-Laien-Kommunikation*, pages 167–185. Frankfurt a/main: P. Lang.

K. Vanopstal and K. Van Wiele. 2007. Incorporation of two terminology projects into a system for information retrieval using nlp for term expansion. In *Proceedings of the International Conference on Language and Health Care*.

J. Vivaldi, L. Marquez, and H. Rodriguez. 2001. Improving term extraction by system combination using boosting. In *Lecture Notes in Computer Science*, pages 515–526.