Classification-Based Scientific Term Detection in Patient Information

Véronique Hoste
Klaar Vanopstal
Els Lefever
Isabelle Delaere

LT3 Language and Translation Technology Team, University College Ghent, Groot-Brittanniëlaan 45, 9000 Gent, Belgium
Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 (S9), 9000 Gent, Belgium

veronique.hoste, klaar.vanopstal, els.lefever, isabelle.delaere@hogent.be

Although intended for the “average layman”, both in terms of readability and contents, the current patient information still contains many scientific terms. Different studies have concluded that the use of scientific terminology is one of the factors, which greatly influences the readability of this patient information. The present study deals with the problem of automatic term recognition of overly scientific terminology as a first step towards the replacement of the recognized scientific terms by their popular counterpart. In order to do so, we experimented with two approaches, a dictionary-based approach and a learning-based approach, which is trained on a rich feature vector. The research was conducted on a bilingual corpus of English and Dutch EPARs (European Public Assessment Report). Our results show that we can extract scientific terms with a high accuracy (>80%, 10% below human performance) for both languages. Furthermore, we show that a lexicon-independent approach, which solely relies on orthographical and morphological information is the most powerful predictor of the scientific character of a given term.

Keywords:
Automatic term extraction, patient information, machine learning

1. Introduction

In a continuously communicating society, there is a growing need for clear, unambiguous communication from public and commercial organizations. This is not only reflected in the increasing number of publications devoted to the (lack of) clarity of government communication (Kimble 2006), and of technical (Zahedi et al. 2001) and medical texts (Rudd et al. 2003), but it has also led to several social and legislative initiatives, such as the European Directive (2001/83/EC) on the readability of the patient information leaflets (PILs) and the concrete elaboration of templates with standard expressions[1]. However, despite the increased scientific and legislative interest in the topic, surveys on PILs (Nink and Schroder 2005, Pandermaat 2008) have shown that respondents often feel distressed by reading the information, or even consider it as fully incomprehensible. This article deals with the automatic identification of one of these sources of distress, namely the use of scientific terminology in patient information. Although intended for the "average layman"[2], both in terms of readability and contents, the current patient information still contains many scientific terms. This might be due to the fact that the public patient information is an adaptation or intergeneric translation (Zethsen 2004) of the scientific leaflet, which is written in expert language with expert terminology.

Our study can be categorized as research into automatic term recognition (ATR), and more specifically as ATR in the biomedical domain. However, whereas most of the work on biomedical ATR focuses on the detection of biomedical terms versus non-terms (Krauthammer and Nenadic 2004, Jacquemin 2001), the present study goes one step further by also distinguishing between different types of terms. This differentiation is crucial because we wish to replace all scientific terms (e.g. *epistaxis*) by their popular counterpart in a next phase (e.g. *nosebleed*). The main aim of this operation is to improve the readability of the patient information. The first step towards the automatic replacement of all scientific terms by a valid popular alternative, is the accurate selection of scientific terminology. In order to do so, we experimented with two approaches, a dictionary-based approach and a learning-based approach, which is trained on a rich feature vector. The research was conducted on a bilingual corpus of English and Dutch EPARs (European Public Assessment Report). Our results show that we can accurately extract scientific terms (F-score: >80%, 10% below human performance) for both languages. Furthermore, we will show that a lexicon-independent approach, which solely relies on orthographical and morphological information is the most powerful predictor of the scientific character of a given term. Our results are restricted to biomedical popular science documents, namely EPAR documents. Given the ultimate goal of this term extraction procedure, viz. the replacement of a scientific term by its popular counterpart, we did not experiment on biomedical scientific documents.

The remainder of this paper is structured as follows. Section 2 motivates the present study and gives an overview of related work on medical terminology and automatic term recognition. Section 3 presents an overview of the English and Dutch corpora being used and discusses the annotation guidelines and the inter-annotator agreement. In Section 4 and 5, we describe the two different approaches, which were taken, viz. lexicon-based versus learning-based term extraction, followed by the experimental setup. Section 6 provides a thorough overview of the feature construction for the learning-based experiments, reports on the results and discusses the main findings of the manual error analysis. Section 7 ends with some concluding remarks and directions for future research.

[1] Other notable initiatives are the "Plain Language" campaign in the UK, the Bill Clinton Memorandum (http://www.plainlanguage.gov) in the US, the "Klare Taal" and "Duidelijke Taal" campaigns in the Netherlands, etc.

[2] As stated in the EMEA report EMEA/126757/2005, 2.0

2. Background and related work

The ability of patients to understand medical information has already been studied extensively. In the domain of oral patient-doctor communication for example, Lerner et al. (2000) conducted a study to determine the emergency department patients' understanding of common medical terms used by health care providers. They asked a balanced group of 249 patients to determine for six pairs of terms whether they had the same meaning or not. Based on the observation that the mean number of correct responses was 2.8 out of 6, they concluded that medical terminology is often poorly understood. Some examples: the percentages of patients that did not recognize analogous terms was 79% for *bleeding* versus *hemorrhage*, 78% for *broken* versus *fractured bone*, 74% for *heart attack* versus *myocardial infarction*. The authors found that especially young, urban, poorly educated patients had a poor understanding of medical terminology. DiFlorio (1991) came to a similar conclusion in a study on the understanding of terms related to the care of newborn babies. In studies with diabetic patients Aufseesser et al. (1995) show similar low understanding.

If we focus on written patient information, similar conclusions have been drawn: despite the legislative efforts, patients still have difficulty understanding the information. In 2005, a survey of the scientific institute of the German AOK (Allgemeine Ortskrankenkassen), Nink and Schroder (2005) revealed that, whereas the majority of the respondents read the leaflet and consider it as an important source of information, one third of the respondents still feels distressed by reading the leaflet. 28% even admit not having taken the drug because of the package insert; 20% consider it as fully incomprehensible. Similar conclusions were drawn for Dutch by Pandermaat (2008). Both studies conclude that the use of scientific terminology is one of the factors which greatly influences the readability of this patient information.

The present study deals with the problem of automatic recognition of overly scientific terminology as a first step towards the replacement of the recognized scientific terms by their popular counterparts. Automatic term recognition (ATR) is crucial in many domains of (computational) linguistics, including automatic translation, text indexing, the automatic construction and enhancement of lexical knowledge bases, etc. In the research on automatic term extraction, two different directions have mainly been taken. On the one hand, the linguistic-based or rule-based approaches, as proposed by Dagan and Church (1994), Ananiadou (1994), Fukuda et al. (1998) and others, 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 extract terms using measures of “unithood” and/or “termhood” to detect candidate terms. Unithood indicates the collocation strength of the units of a term and has been measured by metrics such as mutual information and log-likelihood (Cohen 1995, Fahmi et al. 2007), whereas termhood refers to the association strength of a given term to a domain concept (Medelyan and Witten 2006, Park et al. 2002) (see Section 6 for a thorough description of both measures). 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 meta-learning methods such as boosting (Vivaldi et al. 2001), etc. on feature sets encoding lexical, part-of-speech, 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 a more detailed overview of the field, we refer to Hirshman et al. (2002) and Ananiadou and McNaught (2006). Our approach falls within the category of machine learning approaches, but it differs from the previous described approaches in the specificity of the task, the richness of the feature vector (as shown in Section 6) and the choice of the MBL learning algorithm (which has shown to be quite robust in case of highly skewed data sets).

3. Corpus construction

For both English and Dutch, we collected a parallel corpus of 317 EPAR summaries for the public[3]. EPAR stands for "European Public Assessment Report"; it is a text which is prepared at the end of every centralized evaluation process to provide a summary of the grounds for the opinion in favor of a marketing authorization as taken by the Committee for Human Medicinal Products. The European Medicines Agency, EMEA, makes these EPARs available to the public after deletion of commercially confidential information. Although the EPAR abstracts were originally intended to provide information understandable to the general public, they are often considered as too technical[4].

3.1. Annotation guidelines

For this study, 20 summaries of each language were manually annotated (English: 16,263 tokens; Dutch: 15,938 tokens) by two linguists, who annotated the corpora in parallel and who received free text as input. Their annotation was not restricted to certain morphosyntactic categories. This implies that adjectives, nouns, adverbs, noun phrases, etc. all could receive a label. An overview of the most important morphosyntactic categories that were annotated can be found in Table 1. As annotation environment, Callisto (http://callisto.mitre.org) was used. The main focus in the annotations was to label the scientific terms, which are candidates for replacement by a popular counterpart. The annotators had to differentiate between 4 labels:

- **NamedEntity**: named entities such as *Zostavax*, *Committee for Medicinal Products for Human Use* and *D06BB10*. The named entities were annotated since they often have the same characteristics as scientific terms, whereas they are not candidates for replacement.
- **Scient(ific):** scientific terms such as *neuralgia*, *akathisia*, *gastro-intestinal*.
- **Amb(iguous):** medical terms of which it is assumed that they are widely known and used, e.g. *AIDS*, *antibiotics*.
- **Pop(ular)_Var(iant):** popular variants of scientific terms which are used in the texts. e.g. *People who may be hypersensitive (allergic) to...* or

[3] The EPAR summaries and the annotations will be made publicly available through http://veto.hogent.be/lt3

[4] http://www.emea.europa.eu/pdfs/human/euleg/12675705en.pdf

compared the medicine to a placebo (a dummy vaccine). The terms *allergic* and *dummy vaccine* are labeled as Pop_Var, whereas *hypersensitive* and *placebo* receive a Scient label.

Table 1. Overview of the main morphosyntactic categories that received a label by the annotators

Morphosyntactic categories	*English*	*Dutch*
Nouns	2846	1449
Adjectives	669	385
Verbs	173	101
Conjunctions	78	21
Numerals	50	37
Determiners	45	37
Pronouns	16	6
Prepositions	63	97

The motivation for the different annotations of the medical terms, i.e. **Scient**, **Amb** and **Pop_Var** is the following. Given the objective that the EPARs should be readable by the average layman, the linguists who labeled the data, were asked to give an intuitive label. The objective was to categorize the clear-cut medical slang into the **Scient** category. The terms of which the annotators judged that they were commonly used by many people, were annotated as **Amb**(iguous). Finally, the **Pop_Var** label is mainly relevant in a second experimental phase (not reported in this article) in which the scientific terms have to be replaced by their popular counterpart. All terms, which were not annotated by the annotators will receive a **Pop**(ular) tag for the classification experiments (see for example Table 4).

3.2. Agreement

The annotation agreement between the two annotators was measured by means of the kappa statistic (Carletta 1996). Table 2 gives an overview of the inter-annotator agreement on the different categories and shows similar tendencies for both languages. On both the English and the Dutch data, an agreement score of 0.91 and a kappa score of 0.85 was obtained. Given its ambiguous nature, there is an expected large disagreement on the Amb category. In order to allow for comparison with the results of the classification experiments, we calculated the precision, recall and their weighted F-score for the scientific category by taking the annotation of one annotator as gold standard and the annotation of the other annotator as system output (and vice versa). This led to an F-score of 92.1% for English and 91.8% for Dutch.

Table 2. Contingency table representing the inter-rater agreement for English and Dutch

English	*Scient*	*Amb*	*NamedEntity*	*Pop_Var*	*Total*
Scient	1609	240	9	10	1868
Amb	13	123	0	2	138
NamedEntity	2	0	923	0	925
Pop_Var	2	7	0	191	200
Total	1626	370	932	203	3131

Dutch	*Scient*	*Amb*	*NamedEntity*	*Pop_Var*	*Total*
Scient	1192	114	4	1	1311
Amb	66	69	1	1	186
NamedEntity	21	0	858	0	879
Pop_Var	7	3	0	132	142
Total	1286	186	863	134	2469

For the experiments, the 2 annotated versions of the EPAR summaries were merged using the following hierarchy: Scient > Amb > Pop_Var > NamedEntity. For example, if one annotator decided in favor of a “Scient” tag, whereas the second annotator chose the “Amb” tag, the “Scient” tag was kept in the unified data set.

4. Lexicon-based baseline

For the detection of the medical terms, we experimented both with a lexicon-based and a learning-based approach. Given its simplicity, we considered the dictionary-based system our baseline system.

4.1. English lexicons

For the lexicon-based approach, we collected a number of external lexicons, starting with the MeSH[5] for English. Additionally, we used the Specialist Lexicon[6] from the UMLS, a lexicon which covers both the English general language and concepts from the field of biomedicine and which was originally designed to support the SPECIALIST Natural Language Processing System and to generate indexes to the Metathesaurus. To further expand our lexicon, we also added data from Merriam-Webster's Medical Dictionary[7]. The combined English sources resulted in a lexicon containing 602,873 general and scientific terms. In order to filter out general vocabulary terms we intersected our combined lexicon with the Celex lexical database (Baayen et al. 1993), which resulted in a lexicon of 573,754 unique scientific terms (single word terms: 58%; multi word terms: 42%).

[5] http://www.nlm.nih.gov/mesh

[6] http://lexsrv3.nlm.nih.gov/SPECIALIST/Projects/lexicon/current/web/release/index.html

[7] http://medical.merriam-webster.com/medical/

4.2. Dutch lexicons

In order to build a lexicon for Dutch, a larger number of sources was needed. For the Dutch version of the MeSH, we relied on a termbase as described by Buysschaert (2006). This translation project focuses mainly on chapters C and E (Diseases and Analytical, Diagnostic and Therapeutic Techniques and Equipment respectively). In addition, we used Dutch lexicons such as Taalvlinder[8], Elseviers Medische Encyclopedie[9], the Wikipedia page Gezondheid van A tot Z[10] and the Dutch entries from the Medical Dictionary for Regulatory Activities, MedDRA[11]. We furthermore extracted terminological information from online sources such as:

- Patients' associations e.g.: CMP Vlaanderen[12] & Dystrofie[13]
- Online dictionaries e.g.: Maranje[14]
- Specific websites e.g.: DokterDokter.nl[15]

These sources resulted in a lexicon of 264,778 Dutch terms. Intersecting this lexicon with Celex resulted in a list containing 257,674 scientific terms for Dutch (single word terms: 42%; multi word terms: 42%).

4.3. Experimental setup and results

This external lexical information was the basis for three matching strategies, the results of which are displayed in Table 3. Three measures were used throughout this paper to evaluate the results: Precision, Recall and F-score (Jurafsky and Martin 2009).

$$Precision = \frac{Number\ of\ correctly\ extracted\ terms\ by\ the\ system}{Total\ number\ of\ extracted\ terms\ by\ the\ system}$$

$$Recall = \frac{Number\ of\ correctly\ extracted\ terms\ by\ the\ system}{Total\ number\ of\ actual\ terms\ in\ the\ text}$$

$$F - score = \frac{2(Precision\ *\ Recall)}{Precision\ +\ Recall}$$

First, we matched every single word in the EPARs with the English and Dutch lexicons, which resulted in a rather low recall score, i.e. 22.4% for English and 27.3% for Dutch. In a second step, we experimented with multiword terms and matched every n-gram (up to five) with our lexicon. This multiword term matching, however, was not able to detect new compounds. If, for instance, the terms *lever* (En: *liver*) and *aandoening* (En: *disease*) occurred in the lexicon as separate entries, but *leveraandoening* (En: *liver disease*) did not, the term was lost. This led to a third approach which aims at the detection of new compounds, i.e. fuzzy matching. This fuzzy matching approach matches any entry in our

[8] http://www.ochrid.dds.nl/pages/lijsten/MedAlg.htm
[9] The encyclopaedia can be found at http:/www.kiesbeter.nl/medischeinformatie
[10] http://nl.wikipedia.org/wiki/Gezondheid_van_A_tot_Z
[11] http://www.meddramsso.com
[12] http://www.cmp-vlaanderen.be/content/0311_medwdb.php
[13] http://www.dystrofie.be/termen.html
[14] http://www.medisch.maranje.nl/
[15] http://www.dokterdokter.nl/medisch/begrippen/list/char/A

lexicon combined with another entry. Given the previous example, if both *lever* and *aandoening* occur in the lexicon, *leveraandoening* will be a fuzzy match. The results of our lexicon coverage can be found in Table 3.

Table 3. Coverage of the English and Dutch lexicons

English	***Precision***	***Recall***	***F-score***
Single	78.94	22.39	34.88
Single & Multi	73.43	37.27	49.44
Single, Multi & Fuzzy	47.11	49.02	48.05
Dutch	***Precision***	***Recall***	***F-score***
Single	67.71	27.28	38.89
Single & Multi	51.92	34.08	41.15
Single, Multi & Fuzzy	49.18	37.76	42.72

As expected, lexicon-based term extraction leads to high precision scores at the cost of our recall scores. Table 3 shows a clear difference between the single-word precision scores for both languages, which can be explained through manual analysis of the results. Even though the Dutch lexicon was intersected with Celex in order to filter out the non-scientific words, certain "popular" terms still occur in the lexicon (e.g.: *patient, productie, risico's*). The Dutch lexicon has a 5% higher recall score for the single-word search, which indicates that the entries in this lexicon are more representative of the language used in the EPARs. When we combine the single-word search with the multiword search, precision drops both for English and for Dutch. An explanation can be found in the fact that both lexicons contain multiwords without a scientific character. Using the fuzzy-match approach, higher recall scores were obtained. Precision, however, went down for both languages, which was to be expected given the "fuzzy" character of this approach.

5. Learning-based term extraction

As an alternative to the lexicon-based approach, we experimented with a machine learning based approach to term extraction, which not only exploits lexical information, but also takes into account a rich feature vector incorporating lexical, morphological and statistical information, local-context, etc.

5.1. Memory-based learning

Given the skewedness of the data sets (e.g. English: 2216 scientific tokens in a data set of 16,263 tokens), we experimented with a memory-based learning approach, which in earlier experiments has shown to be quite robust to this data set skewedness (Daelemans et al. 2003a). The approach is based on the memory-based reasoning (Stanfill and Waltz 1986) and case-based reasoning schemes (Riesbeck and Schank 1989, Kolodner 1993), which state that performance in real-world tasks is based on remembering past events rather than creating rules or generalizations. MBL keeps all training data in memory and at classification time, a previously unseen test example is presented to the system and its similarity to all examples in memory is computed using a similarity metric. The class of the most similar example(s) is then used as a prediction for the test instance. This strategy is often referred to as "lazy" (Aha 1997) learning. This storage of all training instances in memory during learning without abstracting and without eliminating noise or exceptions is the distinguishing feature of memory-based learning (MBL) in contrast with minimal-description-length-driven or "eager" ML algorithms (e.g. decision trees, rules and decision lists). *Rule induction*, which can be described as an eager learning approach, compresses the training material by extracting a limited number of rules.

For our experiments, we used the memory-based learning algorithms implemented in TIMBL (Daelemans et al. 2002), which is a fast, decision-tree-based implementation of k-nearest neighbour classification. 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: $(x_1, y_1), (x_2, y_2), (x_3, y_3) \dots (x_n, y_n)$, 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.

- **A distance metric**: When presenting a new instance for classification to the MBL learner, the learner looks in its memory in order to find all instances whose input attributes are similar to the newly presented test instance. In order to do that, we have to define what is meant by similar. In other words, we need to define a *distance metric* that defines how far x_q and x_i are. In order to measure this distance, we calculate the number of matching and mismatching features in two instances (for a description of the feature construction, we refer to Section 6). The distance between x_q and x_i is simply the sum of the differences or distance δ between the *n* features:

  $$Delta(x_q, x_i) = \sum_{i=1}^{n} \delta(x_{qi}, x_{ii})$$

  where:

  $$\delta(x_{qi}, x_{ii}) \;=\; 0 \; if \; x_{qi} \;=\; x_{ii}$$
  $$\delta(x_{qi}, x_{ii}) \;=\; 1 \; if \; x_{qi} \neq x_{ii}$$

This means that all features are considered equally important. This is the approach taken in the IB1 algorithm from Aha et al. (1991). However, IB1 does not solve the problem of modeling the difference in relevance of the various features. In most cases some features will be more informative for the prediction of the class label than others. Therefore, some type of feature weighting is required in which features, which contribute most to a correct classification are given a higher weight than less informative features. For our experiments, we used the default feature weighting method in TIMBL, **gain ratio weighting**, which is a normalized version of the information gain (Quinlan, 1993) weighting method. The information gain of a feature *i* is calculated as follows. Assume we have *C*, the set of

$$H(C) = -\sum_{c \in C} P(c) \log_2 P(c)$$

class labels and V_i, the set of feature values for feature *i*. With this information, we can calculate the database information entropy. The probabilities are estimated from the relative frequencies in the training set.

The information gain of feature *i* is then measured by calculating the difference in entropy between the situations with and without the information about the values of the feature:

$$w_i = H(C) - \sum_{v \in V_i} P(v) \times H(C|v)$$

Gain ratio, is a normalized version of information gain. It is information gain divided by split info *si(i)*, the entropy of the feature values. This is just the entropy of the database restricted to a single feature.

$$w_i = \frac{H(C) - \sum_{v \in V_i} P(v) \times H(C|v)}{si(i)}$$

$$si(i) = -\sum_{v \in V_i} P(v) \log_2 P(v)$$

- **The nearest neighbours**: The nearest neighbours are the instances in memory, which are near to the test item to be classified and the classification of these nearest neighbours is used as classification for the new test instance. The number of nearest neighbours is expressed by *k*. In the original *k*-nearest neighbours algorithm (Cover and Hart, 1967), the *k* closest training examples are taken and the test instance receives the classification of the most common category among these nearest neighbours. In case of continuous feature vectors, Euclidean distance is used to calculate the similarity of two instances. In this case, it rarely happens that two nearest neighbours have the same distance. In case of discrete and symbolic features, however, for which the distance between two values is 0 if they are the same and 1 if different (the above described overlap measure), this occurs regularly. Therefore, in the TIMBL implementation of IB1, *k* refers to the number of nearest distances. For our experiments we used the default *k=1* value. This means that the instances with the nearest distance to the test instance are used for classification. In case of multiple instances at the same distance, TIMBL selects the classification with the highest frequency in the class distribution of the *k*-nearest distances set.

- **A model of how to extrapolate from the nearest neighbours**: in our experiments, we used the default method in TIMBL for deciding which will be the class of a new test item: **majority voting**. This means that all nearest neighbours receive equal weight and that the most frequent class in the nearest neighbour set is taken as classification for the new test item.

5.2. Experimental setup

The general setup for the experiments is the following. All experiments are performed using ***k*-fold cross-validation** (Weiss and Kulikowski 1991) on the data set. This means that the full data set is split into *k* subsets. Iteratively, each partition is used as the hold-out test set while the remaining *k-1/k* balance of the data is used for training. For the experiments, *k* was set to twenty (equaling the total number of annotated EPARs) and the partitions were made at the document level.

6. Feature construction

For the learning experiments, we used a supervised learning approach which is trained on a feature vector set incorporating lexical, morphological and statistical information, local-context, etc. which can contribute to the correct detection of scientific terms. We built a feature vector for each token (the so-called "focus word") in the corpus. This implies that -even for the training data- we did not use any a-priori information on the annotated class labels nor on the part-of-speech, etc. of the words in the corpus. The same feature construction procedure was used both for the training and test data. In a classification approach, such as the one we are using for the term extraction experiments, a classifier is trained on a given training corpus and is then applied to the test data, for which it will assign a class to each test instance using the knowledge it inferred from the training data. Having a class distribution in the test data which closely resembles the training data, set class distribution is of utmost importance for classifier performance.

For example, the training sentence *"The active substance of Abilify is aripiprazole, a quinolinone derivative"*, will be converted to the instances (one for each token) represented in Table 4.

Table 4. Tokens for which a feature vector is constructed, followed by their classification

Focus token	***Class***
The	Pop
active	Scient
substance	Scient
of	Pop
Abilify	NamedEntity
is	Pop
aripiprazole	Scient
,	Pop
a	Pop
quinolinone	Scient
derivative	Scient
.	Pop

The multi word units (MWU) in the learning experiments can be identified by merging feature vectors containing similar classifications: two or more consecutive words, which are classified as “Scient” can be considered a MWU.

6.1. Local context

We included word-form, lemma and part-of-speech information of three words to the left and three words to the right of the focus word. In order to obtain this information, the two corpora were preprocessed by means of a shallow parser (as shown in the example sentence below). The following preprocessing steps were taken both for English and for Dutch. 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. 1996, Daelemans et al. 2003), which was trained on text from the Wall Street Journal corpus in the Penn Treebank, the Brown corpus (Kucera and Francis 1967) and the Air Travel Information System (ATIS) corpus (Hemphill et al. 1990). During text chunking syntactically related words were combined into non-overlapping phrases (represented by square brackets in the example below). Although the chunker provided different types of phrases, we were mainly interested in the NP chunks. These NP chunks are base NPs which contain a head, optionally preceded by premodifiers, such as determiners and adjectives. Postmodifiers are not part of the noun phrase. Part-of-speech tagging and text chunking for Dutch was again performed by the memory-based tagger MBT, this time trained on the Spoken Dutch Corpus (CGN)[16].

[The\DT\The active\JJ\active substance\NN\substance]
of\IN\of [Abilify\NNP\Abilify] is\VBZ\be [aripiprazole\JJ\aripiprazole] ,\,\,
[a\DT\a quinolinone\JJ\quinolinone derivative\NN\derivative] .\.\.

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

6.2. Lexical information

We used the external lexicons to build lexical, binary features using three approaches: single-word matches, multi-word matches and fuzzy-word matches. These approaches and the contents of the lexicons we used are discussed in detail in Section 4.3. In the learning experiments, the matches are just incorporated as features, which implies that a single word match binary feature value of "1" does not necessarily lead to a scientific classification of a given focus word.

As an alternative to integrating dictionary-based lexical information, 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) tried to determine the weight of a word (in a collection of documents) by calculating TF-IDF scores, whereas other researchers, among others Dunning (1993), Rayson and Garside (2000) and Ferreira da Silva et al. (1999), 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) were explored to find lexicon-specific terms.

In order to add corpus-specific lexical information to our feature set, we applied two types of statistical filters on the data (Kageura and Umino 1996):

1. Filters that measure the **Termhood** (Drouin, 2006) or "degree to which a linguistic unit is related to domain-specific context": TF-IDF and Log-Likelihood filters (Section 6.3.)
2. Filters that measure the **Unithood** or "degree of strength or stability of syntagmatic combinations or collocations": Mutual Expectation measure (Section 6.4.)

6.3. Termhood

- **TF-IDF** (term frequency inverse document frequency) is widely used in Information Retrieval to isolate useful keywords in document collections. TF-IDF (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. To calculate the IDF, we enlarged the EPAR corpus with all written and spoken documents of the BNC[17] corpus for English and the TNC[18] for Dutch. Calculating TF-IDF on the EPAR terms should enable us to extract lexicon-specific scientific terms that have much lower frequencies in balanced reference corpora such as the BNC or the TNC.

[17] http://www.natcorp.ox.ac.uk/

[18] http://www.vf.utwente.nl/~druid/TwNC/TwNC-main.html

Given a document collection *D*, a word *w*, and an individual document *d in D*,

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

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).

We used these TF-IDF weighted terms to construct two separate features for our feature vector: one that takes into account the TF-IDF value itself and another one that defines the threshold to perform the intersection with the medical lexicon. In order to determine the TF-IDF threshold for scientific terms, we performed 20-fold cross-validation on the labeled EPAR corpus. First, we calculated the average TF-IDF value of terms that have been manually labeled as being scientific; in order to do so, we ignored the 5% highest and lowest values in all 20 training runs. This led to the selection of 1.05 as threshold, which was used to create a binary TF-IDF feature. This threshold value was also used to rebalance the highly skewed data set (as we will explain in the next section). We measured both the percentage of correctly labeled terms (scientific terms having a TF-IDF value above the threshold and popular terms having a TF-IDF value below the threshold) as well as precision and recall for the scientific terms.

- 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 the frequency lists. This is done by constructing a contingency table as is shown in Table 5, where *c* represents the number of words in the first corpus, while *d* corresponds to the number of words in the second corpus. The values *a* and *b* are called the observed values (*O*).

Table 5. Contingency table to calculate Log-Likelihood

	First Corpus	*Second Corpus*	*Total*
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	c	c+d

In the formula below, *N* corresponds to the total number of words in the corpus, *i* corresponds to the single words, whereas the "observed values" O_i correspond to the real frequency of a single word *i* in the corpus. For each word *i*, the observed value O_i is used to calculate the expected value E_i according to the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

Applying this formula to our contingency table (with $N_1 = c$ and $N_2 = d$) results in:

$$E_1 = c * (a + b)/(c + d)$$
$$E_2 = d * (a + b)/(c + d)$$

We then used the resulting Expected values for the calculation of the Log-Likelihood:

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

which equates to:

$$LL = 2 * ((a * log(\frac{a}{E_1})) + (b * log(\frac{b}{E_2})))$$

The formula for the calculation of both the expected values (E) and the Log-Likelihood have been described in detail by Rayson and Garside (2000). 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 frequencies). The log-likelihood information was integrated as a binary feature. Terms with log-likelihood value above a predefined threshold and with BNC frequency below a predefined threshold were set to 1, the others were set to 0. Both thresholds were validated on the EPAR corpus using 20-fold cross-validation.

6.4. Unithood

As a measure of unithood, we calculated the Mutual Expectation values for the 1- to 8-grams. Dias and Kaalep (2003) developed the Mutual Expectation (ME) Measure to test the cohesiveness between words in a multiword term, i.e. the group of words forming a multiword should occur together more frequently than expected by chance. In order to calculate the Mutual Expectation values, the n-gram frequencies (up to 8-grams) are calculated based on the EPAR corpus and used to derive the Normalized Expectation (NE) values for all multiword terms, as specified by following formula:

$$NE = \frac{prob(n - gram)}{\frac{1}{n}\sum prob(n - 1 - grams)}$$

This Normalized Expectation expresses the cost, in terms of cohesiveness, of losing one component of the n-gram. In case the cohesiveness of the multiword is very high, the frequency of the n-gram minus one component is expected to be lower, and the resulting Normalized Expectation value will be high again. For example, if we compare the two bigrams “protease inhibitors” (ME: 9.3) and “the wart” (ME: 0.0001), it is already intuitively clear that the resulting unigram “the” when deleting the last word of the bigram ("wart") will be much more frequent than the resulting unigram "protease". As simple n-gram frequency appears to be a valid criterion for multiword term identification (Daille 1995), the final Mutual Expectation values are obtained by multiplying the Normalized Expectation and the relative frequency of the multiword. We calculated Mutual Expectation values for all English and Dutch multiword terms in our EPAR corpus and performed an additional filtering on the list of multiwords. We only included multiwords in case:

1. one of the words of the multiword appeared in our scientific lexicon
2. the multiword formed a valid syntactical chunk (noun phrase, prepositional phrase or a combination of both)
3. the multiword contained at least one content (i.e non-grammatical) word

We included the Mutual Expectation information as a binary feature: in case the value is higher than a predefined threshold (0.1 in our case) and the target word complies with the three conditions mentioned above, the feature is set to "yes", otherwise the feature is set to "no". In order to define the threshold, we calculated ME values for all multiwords occurring in our full EPAR corpus, applied the three filtering criteria and sorted the list according to descending ME values.

6.5 Morphological information

In addition to the features above, cognates may also be an indication of termhood. Other useful morphological features are Greek and Latin affixes.

- **Cognate information**:
  Cognate matching has been successfully used to find correspondence points for the alignment of parallel texts (Kondrak 2003, Simard et al. 1992 and Ribeiro et al. 2001) and to extract terms from bitexts (Alegria et al. 2006). In a linguistic context, cognates are words, which share the same origin and have similar orthography. In computational linguistics, however, cognates are defined as words, which have similar orthographic and identical semantic properties (Melamed 1999). Previous work on cognate detection has been focused both on orthographic evidence, such as the Levenshtein distance, Dice's coefficient, Longest Common Subsequence Ratio, and on semantic evidence (Mulloni et al. 2007). A combinatory method using both orthographic and semantic evidence has been discussed by Nakov et al. (2007) and Mitkov et al. (2007). Semantic evidence is particularly valuable for the detection of false cognates, especially from comparable corpora. However, the problem of false friends is negligible in this research: the use of word-aligned bitexts reduces the chance of finding false friends to a minimum.

  We composed a list of cognates, which was incorporated into the system as a binary feature. Two sources were used for the compilation of this list: MeSH translations (Buysschaert 2006) and 20 English and Dutch EPARs. We split up all multiword terms in the English-Dutch MeSH list and extended this list with the resulting single-word terms. Subsequently, we calculated the Longest Common Subsequence Ratio (Hirschberg 1977) for each of these terms, which involves finding the longest subsequence common to the pair of sequences. We then filtered out those terms with a substring overlap of less than four characters. This resulted in a list of 6,495 unique English and 5,646 unique Dutch cognate terms. The difference in number between English and Dutch can be explained by the fact that some English synonyms are translated by the same Dutch term.

To further extend this list, we manually aligned 20 English and Dutch EPARs on the sentence level, and used the output to automatically align them on word level, using the Perl implementation of IBM Model One that is part of the Microsoft Bilingual Sentence Aligner (Moore 2002). The candidate terms were tokenized and a POS filtering provided us with a list containing mainly nouns and adjectives. Subsequently, the Longest Common Subsequence Ratio was calculated for each word pair and again only those terms with a substring overlap of more than four characters were taken into account. This way, 45 English and 54 Dutch cognates were added to the list. Finally, as already mentioned, this information was integrated as a binary feature.

- **Affix information**: Medical terminology has the specificity to use abbreviations, acronyms and Latin terms (Surjan and Heja 2003). Affixation and (semi-)neoclassical compounding have proven 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). 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, four additional features were deduced: the presence of a prefix (e.g. *condyl*ar canal), the presence of a suffix (e.g. ir*itis*), the presence of both prefix and suffix in one term (e.g. *dys*lipid*aemia*) and the presence of a confix in the centre of a term (e.g. anti*hist*amines).

6.6. Orthographic features and trigrams

Two orthographic features were used as an indication of whether a given word is a scientific term or not. The first orthographic feature verifies whether a given word consists of or contains numeric symbols, a characteristic that may indicate that it is indeed a scientific term (e.g. b2-microglobulin). The second orthographic feature detects whether a given word contains multiple capital letters, which could indicate an abbreviation or an acronym (e.g. HIV).

Furthermore, we included two trigram features, which represent the initial and final trigram of a given word.

6.7. Term patterns

A short analysis of the local contexts of terms labeled as scientific in the annotated EPAR corpus showed that several term patterns can be detected. Some phrases, denoted by Pearson (1996) as hinges, may signal the presence of a term, for example "is referred to as", "denotes", "is defined as", "is called", "known as" etc. In total, we gathered 14 left context-patterns for English and 4 for Dutch for our experiments. We built a binary feature to verify whether a given focus word was preceded by one of these patterns or not.

6.8. Contribution of the different types of information sources

Table 6. 20-fold cross-validation results on the English EPAR data set with the complete feature vector. Contribution of the different types of feature information.

	All classes	***Scientific class***		
ENGLISH	Accuracy	Precision	Recall	F-score
		COMPLETE SYSTEM		
All features	92.36	85.59	77.75	81.48
	BASELINE SYSTEMS USING GROUPS OF FEATURES			
Local context	89.61	76.64	66.92	71.45
Local cont. - focus	84.97	62.06	58.75	60.36
TF-IDF, LL and ME	77.11	50.42	40.39	44.85
Lexical	78.71	63.51	38.18	47.69
TF-IDF + lexical	80.23	73.81	43.50	54.74
Orthographic	88.21	81.24	63.72	71.42
Morphological	78.97	71.89	36.46	48.38
Orthographic and morphological	88.78	84.34	67.33	74.88

Table 7. 20-fold cross-validation results on the Dutch EPAR data set with the complete feature vector. Contribution of the different types of feature information.

	All classes	***Scientific class***		
DUTCH	Accuracy	Precision	Recall	F-score
		COMPLETE SYSTEM		
All features	93.92	86.01	76.48	80.97
	BASELINE SYSTEMS USING GROUPS OF FEATURES			
Local context	91.07	71.23	63.39	67.08
Local cont. - focus	86.77	53.17	51.92	52.54
TF-IDF, LL and ME	82.56	49.86	23.97	32.38
Lexical	83.20	62.95	34.20	44.32
TF-IDF + lexical	85.35	76.39	38.57	51.26

Orthographic	91.07	82.55	64.10	72.17
Morphological	84.37	81.59	34.07	48.07
Orthographic and morphological	90.95	80.58	65.67	72.36

Tables 6 and 7 show 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 F-score on the scientific terms, our category of interest. Overall, we can observe an F-score of 81% for the detection of scientific terms both in the English and the Dutch EPARs. Furthermore, we can observe that the precision scores are consistently higher than the recall scores for both languages.

Considering the contribution of the different feature types, some observations can be made. The combination of the orthographic and morphological features gave the best results concerning F-score both for English and for Dutch. It appears that the information about prefixes, suffixes, trigrams, capitalization and word-internal numbers highly influences our system. It is also remarkable how much the system benefits from the local context information, which includes word form, lemma and part-of-speech information. There is a significant difference between the results, which include the focus word and its additional information and the results which do not. The combination of TF-IDF, Log-Likelihood and external lexical information results in high precision scores (respectively 73.8% and 76.4%), but rather low recall (43.5% and 38.6%), which was to be expected given the low coverage results of our lexicons.

Table 8. Confusion matrix showing the number of words per error class for English and Dutch. Column labels are referring to manual annotation whereas row labels refer to system output.

ENGLISH	Scient	Amb	NamedEntity	Pop_Var	Popular	Total
Scient	**1723**	55	37	22	176	2013
Amb	74	**277**	0	5	73	429
NamedEntity	62	1	**991**	1	16	1071
Pop_Var	47	8	1	**113**	85	254
Popular	310	105	31	134	**11916**	12496
Total	2216	446	1060	275	12266	**16263**
DUTCH	Scient	Amb	NamedEntity	Pop_Var	Popular	Total
Scient	**1179**	46	13	12	121	1371
Amb	62	**151**	1	6	55	275
NamedEntity	46	5	**862**	1	36	950
Pop_Var	14	12	1	**80**	89	196

Popular	238	69	48	93	**12698**	13146
Total	1539	283	925	192	1299	**15938**

The confusion matrix in Table 8 shows the number of entries per error class for English and for Dutch. Both languages seem to share “problem areas”, i.e. classes that have a higher entry number. The most problematic classes are the following:

- **Scientific terms being predicted as popular terms**: this error class is the most problematic one, as our final goal is to develop a system which automatically replaces scientific terms by their popular counterparts.

  Analysis of the entries in this class has shown us the following: words are sometimes labeled correctly as scientific but received the popular label on other occasions (e.g. *replacement* insulin). When we looked in more detail at these examples, it appeared that the terms were not always given the same label by the annotators. It also seems that when the focus word appears at the beginning or the end of a sentence, and therefore lacks local context, the error rates go up. The coverage of our lexicon is also a factor in this error class as many terms do not occur in it, which means we will have to expand our lexicon. Detailed analysis showed that the occurrence of a product name in the local context to the left of the focus word has a negative influence for this particular class, i.e. scientific terms being labeled as popular. Another type of words that can be listed in this error class are those that are only rarely labeled as scientific (e.g. *but, patients, the*) and that appear in this error class when they are part of a multi-word term. On certain occasions (e.g. an *inherited* disease), the given focus word is detected as part of a larger multi-word term. However, the weight of this sole feature is often not substantial enough for the system to label the word as scientific.

- **Popular terms being predicted as scientific terms**: after thorough manual analysis of the entries in this class some characteristics can be detected. It appears that, if the local context of the focus word contains a bracket, a colon, capital letters or numeric symbols, the focus word easily receives the scientific label. In this error class, scientific terms preceding or following the focus word and the inconsistent labeling by the annotators are also two of the main factors that cause the system to label the focus word erroneously. An examples of this error class is “ketoconazole *or* itrconazole” where the focus word *or* is preceded and followed by a scientific term.

- **Popular variants being predicted as popular words**: this is an obvious error class, as popular variants are in fact popular words. The difference between these two classes lies in the fact that a popular word can only be a so-called popular variant if it occurs near a scientific term. However, our system is able to detect about 42% of the popular variants correctly, both for English and for Dutch. Therefore, it remains interesting to look at this error class in detail. One of the factors that seem to influence the system in a negative way is the absence of one of the brackets that normally appear in the popular variant's local context (e.g. *asthenia (weakness)*). These brackets may not appear in the local context for a number of reasons: the focus word is centered in a very large string of popular words (e.g. a definition), the focus word appears at the beginning of a larger string of popular variants and therefore misses the second bracket in the local context, no brackets are used etc. Finally, as popular variants are actually a subclass of the popular words, the focus word may have been annotated as "popular" on numerous occasions, whereas the "popular variant" label is only given under certain circumstances (i.e. if it appears near a scientific term).

7. Concluding remarks

In this paper, we investigated the use of a dictionary-based and a machine learning approach to scientific term detection in patient information. The learning approach not only exploits lexical information, but also takes into account a rich feature vector incorporating lexical, morphological and statistical information, local-context, etc. We showed an F-score of above 80% for the prediction of scientific terms in an English and a Dutch EPAR corpus. We expect these results to improve when increasing the size of the data set.

As a next step, we plan to use genetic algorithms to obtain the optimal feature selection for our classification task. We plan to automatically replace the detected scientific terms by their popular counterparts. In case no popular counterpart is available, a definition will be proposed.

We have made some preliminary experiments with the semi-automatic replacement of scientific terms by their popular counterpart/definition. In these experiments, the authors of patient leaflets used an authoring environment in which (1) all terms were categorized following the methodology described in this article and in which (2) an alternative (popular counterpart, definition) was proposed for the scientific terms. A preliminary readability test shows that this indeed leads to the improved readability of patient information.

Acknowledgments

The research of Isabelle Delaere is conducted within the framework of the IWT-TETRA funded ABOP (Automatic Optimizer of Patient Information) project (IWT-code: 70103).

References

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

Aha, D. 1997. "Lazy learning: Special issue editorial." *Artificial Intelligence Review* 11, 7-11.

Alegria, I., Gurrutxaga, A., Saralegi, X. and Ugartetxea, S. 2006. "Elexbi, a basic tool for bilingual term extraction from spanish-basque parallel corpora." In *Proceedings of the 12th EURALEX International Congress*, 159-165.

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

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

Aufseesser, M., Lacroix, A., Binyet, S. and Assal, J. P. 1995. “Diabetic retinopathy. interpretation of medical terms by patients.” *J Fr Ophtalmol* 18 (1), 27-32.

Baayen, R., Piepenbrock, R. and van Rijn, H. 1993. *The celex lexical data base on cd-rom*.

Banay, G. 1948. “An introduction to medical terminology in greek and latin derivations.” *Bulletin of the Medical Library Association* 1 (36), 1-27.

Berger, A., Caruana, R., Cohn, D., Freitag, D. and Mittal, V. 2000. “Bridging the lexical chasm: Statistical approaches to answer finding.” In *Proceedings of the International Conference on Research and Development in Information Retrieval,* 192-199.

Buysschaert, J. 2006. “The development of a mesh-based biomedical termbase at hogeschool Gent.” In *Proceedings of the LREC 2006 Satellite Workshop W08. Acquiring and representing multilingual, specialized lexicons: the case of biomedicine*, 39-43.

Carletta, J. C. 1996. “Assessing agreement on classification tasks: the kappa statistic.” *Computational Linguistics* 22 (2), 249-254.

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

Cohen, W. W. 1995. “Fast effective rule induction.” In *Proceedings of the 12th International Conference on Machine Learning (ICML-1995)*, 115-123.

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

Cover, T. and Hart, P. 1967. “Nearest neighbour pattern classification.” *Institute of Electrical and Electronics Engineers Transactions on Information Theory* 13, 1-27.

Daelemans, W., Zavrel, J., Berck, P. and Gillis, S. 1996. “Mbt: A memory-based part of speech tagger generator.” In *Proceedings of the 4th ACL/SIGDAT Workshop on Very Large Corpora*, 14-27.

Daelemans, W., Zavrel, J., van der Sloot, K. and van den Bosch, A., 2002. “Timbl: Tilburg memory-based learner, version 4.3, reference guide.” *Tech. Rep. ILK Technical Report - ILK 02-10*, Tilburg University.

Daelemans, W., Hoste, V., De Meulder, F. and Naudts, B., 2003a. “Combined optimization of feature selection and algorithm parameter interaction in machine learning of language.” In *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*, 84-95.

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

Dagan, I. and Church, K. 1994. “Termight: identifying and translating technical terminology.” In *Proceedings of Applied Language Processing*, 34-40.

Daille, B. 1995. “Combined approach for terminology extraction: lexical statistics and linguistic filtering.” *Tech. Rep. 5*, Lancaster University: UCREL.

Dias, G. and Kaalep, H. 2003. “Automatic extraction of multiword units for estonian: Phrasal verbs.” *Languages in Development* 41, 81-91.

DiFlorio, I. 1991. “Mothers' comprehension of terminology associated with the care of a newborn baby.” *Pediatr Nurs* 17 (2), 193-196.

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

Drouin, P. 2006. “Termhood experiments: quantifying the relevance of candidate terms.” *Modern Approaches to Terminological Theories and Applications (Picht, H. ed.), Linguistic Insights* 36, 375-391.

Fahmi, I., Bouma, G. and van der Plas, L. 2007. “Using multilingual terms for biomedical term extraction.” In *Proceedings of the Workshop on Acquisition and Management of Multilingual Lexicons*, 27-34.

Ferreira da Silva, J., Dias, G., Guilloré, S and Pereira Lopes, J.G. 1999. “Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units.” In *Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, 113-132.

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

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

Hemphill, C., Godfrey, J. and Doddington, G. 1990. “The atis spoken language system pilot corpus.” In *Proceedings of the DARPA Speech and Natural Language Workshop*, 96-101.

Hirschberg, D. S., 1977. “Algorithms for the longest common subsequence problem.” *J. ACM*, 24 (4), 664-675.

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

Jacquemin, C. 2001. *Spotting and Discovering Terms Through Natural Language Processing*. Cambridge, MA: MIT Press.

Jurafsky, D. and Martin, J. 2009. *Speech and Language Processing*. New Jersey, Prentice Hall.

Kageura, K. and Umino, B. 1996. “Methods of automatic term recognition: a review.” *Terminology* 3 (2), 259-289.

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

Kilgarriff, A. 2001. “Comparing corpora.” *International Journal of Corpus Linguistics* 6 (1), 1-37.

Kimble, J. 2006. *Lifting the Fog of Legalese*. Carolina Academic Press.

Kolodner, J. 1993. *Case-based reasoning*. Morgan Kaufmann, San Mateo, CA.

Kondrak, G. 2003. “Cognates can improve statistical translation models.” In *Proceedings of HLT-NAACL 2003*, 46-48.

Krauthammer, M. and Nenadic, G. 2004. “Term identification in the biomedical literature.” *Biomedical Informatics* 37 (6), 512-526.

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

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

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

Lerner, E. B., Jehle, D. V., Janicke, D. M. and Mosati, R. M., 2000. “Medical communication: do our patients understand?” *Am J Energ Med* 18 (7), 764-766.

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

Medelyan, O. and Witten, I. H. 2006. “Sense disambiguation workshop: Recent successes and thesaurus based automatic keyphrase indexing. “ In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 296-297.

Melamed, D. 1999. “Bitext maps and alignment via pattern recognition.“ *Computational Linguistics* 25 (1), 107-130.

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

Mitkov, R., Pekar, V., Blagoev, D. and Mulloni, A. 2007. “Methods for extracting and classifying pairs of cognates and false friends.“ *Machine Translation* 21 (1), 29-53.

Moore, R. C. 2002. “Fast and accurate sentence alignment of bilingual corpora.“ In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, 135-144.

Mulloni, A., Pekar, V., Mitkov, R. and Blagoev, D. 2007. “Semantic evidence for automatic identification of cognates.“ In *Proceedings of A Workshop on Acquisition and Management of Multilingual Lexicons*, 49-54.

Nakov, S., Nakov, P. and Paskaleva, E. 2007. “Cognate or false friend? ask the web!“ In *Proceedings of A Workshop on Acquisition and Management of Multilingual Lexicons*, 55-62.

Nink, K. and Schroder, H. 2005. “Zu risiken und nebenwirkungen: Lesen sie die packungsbeilage?“ *Tech. rep. Wissenschaftliches Institut der AOK (WIdO), WIdO-Materialien* Bd. 53.

Pander Maat, H. 2008. “Hoe (on)leesbaar zijn geneesmiddelenbijsluiters? Een test van drie veel gebruikte bijsluiters“. http://www.let.uu.nl/Henk.PanderMaat/personal/begrijpelijkheid_doc/rapport%2015-4.pdf

Park, Y., Byrd, R. J. and Boguraev, B. 2002. “Automatic glossary extraction: Beyond terminology identification.“ In *Proceedings of the 19th International Conference on Computational Linguistics*, 1-7.

Pearson, J. 1996. “Strategies for identifying terms in specialised texts.“ *Tech. Rep. 16*, Irish Association for Applied Linguistics.

Quinlan, J. 1993. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo, CA.

Rayson, P. and Garside, R. 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)*, 1-6.

Ribeiro, A., Dias, G., Lopes, G. P. and Mexia, J. T. 2001. “Cognates alignment.“ In Maegaard, B. (Ed.) *Proceedings of the Machine Translation Summit VIII (MT Summit VIII)*, Santiago de Compostela, 287-292.

Riesbeck, C. and Schank, R. 1989. *Inside Case-Based Reasoning*. Lawrence Erlbaum Associates, Cambridge, MA.

Rudd, R., Comings, J. and Hyde, J. 2003. “Leave no one behind: Improving health and risk communication through attention to literacy.“ *Journal of Health Communication* 8 (3), 104-115.

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

Simard, M., Foster, G. F. and Isabelle, P. 1992. “Using cognates to align sentences in bilingual corpora.“ In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, 67-81.

Stanfill, C. and Waltz, D. 1986. “Toward memory-based reasoning.“ *Communications of the ACM* 29 (12), 1213-1228.

Surjan, G. and Heja, G. 2003. “About the language of hungarian discharge reports.“ *Stud Health Technol Inform* 22, 869-873.

Vanopstal, K. and Van Wiele, K. 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*.

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

Weiss, S. M. and Kulikowski, C. A. 1991. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, San Mateo, CA.

Zahedi, F. M., Pelt, W. V. V. and Song, J. 2001. “A conceptual framework for international web design.“ *IEEE Transactions on Professional Communication* 44 (2), 83-103.

Zethsen, K. 2004. “Latin-based terms. true or false friends?” *Target* 16 (1), 125-142.