# The Automatic Detection of Scientific Terms in Patient Information

Véronique Hoste, Klaar Vanopstal and Els Lefever
LT3 Language and Translation Technology Team
University College Ghent
Groot-Brittanniëlaan 45
9000 Gent, Belgium
*veronique.hoste@hogent.be*, *klaar.vanopstal@hogent.be*, *els.lefever@hogent.be*

## Abstract

Despite the legislative efforts to improve the readability of patient information, different surveys have shown that respondents still feel distressed by reading the information, or even consider it as fully incomprehensible. This paper deals with one of the sources of distress: the use of scientific terminology in patient information. In order to assess the scale of the problem, we collected a Dutch-English parallel corpus of European Public Assessment Reports (EPARs) which was annotated by 2 annotators. This corpus was used for evaluating and training an automatic approach to scientific term detection. We investigated the use of a lexicon-based and a learning-based approach which only relies on text-internal clues. Finally, both approaches were combined in an optimized hybrid learning-based term extraction experiment. We show that whereas the lexicon-based approach yields high precision scores on the detection of scientific terms, its coverage remains limited. The learning-based approach on the other hand demonstrates an F-score of 80% and remains quite robust despite the highly skewed data set.

## Keywords

Automatic term extraction, scientific terminology, patient leaflets, lexicon-based extraction, machine learning

# 1 Introduction

The pilot study deals with the identification of scientific terminology in patient information such as the patient information leaflet (PIL) or the European Public Assessment Report (EPAR). Previous research [16] has shown that the use of scientific terminology is one of the factors which greatly influences the readability of this patient information. Despite the legislative efforts to improve the readability of the text type, different surveys have indicated that many patients still have difficulty understanding the information. Recently, for example, the scientific institute of the German AOK ("Allgemeine Ortskrankenkasse")[13] conducted a survey on the attitude of their clients toward patient leaflets. Although the results of this survey reveal that the majority of the respondents read the leaflet and also consider it as an important source of information, one third of the respondents still feels distressed by reading the leaflet. 28% even admits not having taken the drug because of the package insert; 20% considers it as fully incomprehensible. Both the leaflet and EPAR suffer from two translation operations[18]: intergeneric translation (translation between genres) and inter-linguistic translation (translation between languages). Inter-generic translation is often problematic since the leaflet for the public is mostly an adaption of the scientific leaflet, which is 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.

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 the research on automatic term extraction, two different directions mainly have been taken: (i) the linguistic-based approaches, such as the term extraction tool developed by Dagan and Church [8], look for specific (mostly language-specific) linguistic structures that match a number of predefined syntactic patterns, whereas (ii) the statistical corpus based approaches (e.g the Term Extractor developed by Pantel and Lin [14]) extract terms using metrics such as mutual information and log-likelihood to measure the information between words. Hybrid approaches combining both linguistic and statistical information have also emerged ([11], [12]). In this paper, we investigate the use of a machine-learning based approach to the specific problem of scientific term detection in patient information. 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.

The remainder of this paper is structured as follows. Section 2 gives an introduction to the specific problem of scientific versus popular terminology in popular patient information and presents an overview of the Dutch and English corpora being used. Section 3 presents two baselines. As a first baseline, we investigate a lexicon-based baseline approach to the problem, based on a combination of medical and general lexical resources. As a second baseline, we experiment with a machine learning approach trained on text-internal features. Section 4 gives an overview of a hybrid machine learning experiment conducted on the data, using both external information sources and intra-textual features. Through the GA-based interleaved optimization of feature selection and parameter settings, we show which information sources contribute most to an accurate detection of scientific terminology. Section 5 continues with the description of the main findings in a manual error analysis on the classifier results. Section 6 concludes this paper.

## 2 The Problem of Scientific Terms in Patient Information

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 such as the following:

> The active substance of Abilify is **aripiprazole**, a **quinolinone derivative**. The primary **pharmacodynamics** of **aripiprazole** suggests that its efficacy is mediated through a combination of partial **agonist** activity at **dopamine D2 receptors** and **serotonin 5-HT1A receptors** and **antagonism** at **serotonin 5-HT2A receptors**.

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, which stands for "European Public Assessment Report", is a text which is prepared at the end of every centralized evaluation process to provide a summary of de grounds for the opinion in favor of a marketing authorization as taken by the Committee for Human Medicinal Products. The EMEA makes these EPARs available to the public after deletion of commercially confidential information. Although these EPAR abstracts were originally intended to provide information understandable to the general public, they suffer from the same shortcomings as the package leaflets which are also often considered as too technical.
But how can we determine in an objective way whether a given term can be considered as scientific or not? Some people are well informed over their illness. Others are less so, maybe due to differences in age, intelligence, social background or just in how they wish to deal with their situation. EMEA (report EMEA/126757/2005,2.0) states that the summaries target the "average layperson", both in terms of readability and contents.

### 2.1 A Dutch and English EPAR corpus

For both Dutch and English, we collected a parallel corpus of 317 EPAR summaries[1]. For this pilot study, 20 summaries of each language were manually annotated (English: 17,511 tokens; Dutch: 17,093 tokens) by two linguists, who annotated the corpora in parallel. As input, they received free text, which was tokenized and provided with lemmatization, POS and chunk information. Tokenization for both languages was performed by a rule-based system using regular expressions. Lemmatization for Dutch was performed by a memory-based lemmatizer trained on a lexicon derived from the Spoken Dutch Corpus (CGN)[2], a 10-million word corpus of spoken Dutch. The English lemmatizer was trained on Celex. Part-of-speech tagging and text chunking were performed by the memory-based tagger MBT[7], which was also trained on the CGN corpus for Dutch and on the Wall Street Journal corpus in the Penn Treebank [10] for English.

The annotators annotated chunks and had to differentiate between the following three labels:

- `scientific`: Terms being labeled as 'scientific' can be real scientific terms (E.g *serotonin*, *schizophrenia*, *akathisia*, *gastro-intestinal*), product names (E.g *ABILIFY*) and their International Nonproprietary Name (E.g *Aripiprazole*) and more technical or specialised terms that might be hard to understand for random users (E.g *derivative*, *intravenously*, *post-menopausal*, *symptomatic*).

- `medium`: The 'medium' label is used for terms that are used with a specific medical meaning (E.g *[renal] compromise*, *depression*, *infection*, *treatment*) or consecutive terms that form together frequently used medical expressions (E.g *psychotic disorders*, *treatment response*, *adverse events*, *weight reduction*). This category was created since we expected that a binary annotation task of distinguishing between scientific and non-scientific terms, would lead to a too polarized view on the data.

- `popular`: 'popular' is considered to be the default label and refers to all general vocabulary terms.

Both annotators also adhered to the following conventions. For the two first categories, they focused on pharmaceutical product names, scientific and technical terms, or general vocabulary terms used in a specific medical context. Tokens referring to place names, company names and arbitrary alphanumerical codes (ATC codes) for pharmaceutical products were by default popular. One might argue that the ATC codes for pharmaceutical products should be labeled as scientific. However, given our ultimate goal of replacing every scientific term by its popular counterpart (if existing), we decided to focus on the terms which come into account for replacement.

The parallel texts show similar tendencies with respect to inter-annotator agreement. On the English data, an agreement score of 0.90 and a kappa score of 0.64 were obtained. For Dutch, the equivalent scores were 0.94 and 0.76. Table 1 gives the contingency table for both

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

[2] http://lands.let.ru.nl/cgn

| `English` | scientific | medium | popular | total |
|---|---|---|---|---|
| scientific | **1284** | 284 | 251 | 1819 |
| medium | 258 | 319 | 693 | 1270 |
| popular | 111 | 146 | 14,165 | 14,422 |
| total | 1653 | 749 | 15,109 | 17,511 |
| `Dutch` | scientific | medium | popular | total |
| scientific | **1423** | 174 | 186 | 1783 |
| medium | 152 | 167 | 228 | 547 |
| popular | 76 | 162 | 14,525 | 14,763 |
| total | 1651 | 503 | 14,939 | 17,093 |

**Table 1:** *Contigency table representing the inter-rater agreement for Dutch and English.*

data sets. It shows two rather imbalanced data sets, with one predominant label covering about 90% of all assigned labels. Since no restrictions were given to the annotators with respect to part of speech category, we can expect that the large majority of tokens will be labeled as popular terms, which was indeed the case. Due to its "being in the middle" position, the annotators expectedly disagreed heavily on the 'medium' category. Since this paper focuses on the use of scientific terms in popular medicinal texts, our category of interest is the cell representing agreement and disagreement on scientific terms. For English, both annotators agreed that 1284 terms, i.e. 7.3% should be labeled as scientific terms. They also agreed on a scientific label for 8.3% of the tokens in the Dutch EPARs. Overall, both annotators give a scientific tag to about 10% of all tokens. For the experiments on scientific term extraction in Sections 3 and 4, we only considered the diagonal cells representing the agreement on the scientific terms.

## 2.2 Inter-language agreement

In addition to inter-annotator agreement, we also measured the inter-language agreement by investigating whether Dutch and English use the same type of term for the same phenomenon. Obviously lacking a one-to-one correspondence and due to the multiple translation shifts between the texts in the two languages, we performed a manual analysis of part of the corpus (2700 words per language) in order to reveal these inter-language labeling differences. In order to avoid interference with the inter-annotator scores, we only considered the labels given by one annotator. As this annotater was a non-native speaker of English, there might be some misperceptions as to the degree of 'scientificness' of certain terms. However, this will be limited to a minority of terms, considering the annotator's academic background in English linguistics. Since the popular terms are of no interest for this task, a total of 461 term pairs were extracted, approximately half of which were labeled unambiguously as scientific. The inter-language agreement on 'medium' terms, however, is much lower.

Some general conclusions can be drawn from the results of this analysis. From the figures in the confusion matrix one can conclude that a comparable number of terms are considered as scientific in both English and Dutch (294 vs. 289). However, a minority of the terms labeled as 'scientific' do not match in both languages. In English, there is a tendency towards the use of abbreviations whereas the Dutch EPARs count more full forms (e.g. *CIU* for *chronische idiopatische urticaria*, *SAR* for *seizoensgebonden allergische rhinitis*, *PAR* for *perennerende allergische rhinitis*). With respect to the differences in labeling between the two languages, some general observations can be made:

|  | scient. | med. | pop. | no trans. | total |
|---|---|---|---|---|---|
| scientific | **237** | 15 | 24 | 18 | 294 |
| medium | 26 | 38 | 48 | 3 | 115 |
| popular | 19 | 23 | - | - | - |
| no trans. | 7 | 3 | - | - |  |
| total | 289 | 79 | - | - | - |

**Table 2:** *Contigency table representing the inter-language agreement for Dutch (vertical) and English (horizontal) scientific and medium terms.*

- Both languages have different term formation patterns: in English, Latin-based terms are much more common than in Dutch. These English Latin-based terms often have no popular equivalent [18], contrary to most Dutch Latin-based terms. *Palate*, for instance, is translated by *palatum*, a Dutch scientific term which has a popular equivalent *gehemelte*. Similarly, *redistribution* is translated by the Dutch term *redistributie*, which has *herverdeling* as a more popular variant. Loan words form another category of terms which may have a more scientific connotation to speakers of Dutch. *Assembly*, for example, is translated with the morphologically similar scientific term *assemblage*, which has a popular counterpart in Dutch, namely *samenvoeging* or *verzameling*.

- Different labeling can also be explained by translation shifts which cause switches in register between English and Dutch. For example, *nasal discharge*, a more scientific term for *runny nose*, is translated by a popular term *loopneus*. Similarly, the term *agents* pertains to a scientific domain, mainly that of chemistry, and is translated into Dutch as *middelen*, which is a general-language word. Moreover, translation shifts, or more specifically class shifts in which adverbs may be translated by adjectives or adjectives by nouns etc, make the detection of scientific terms more difficult (e.g. *were actively symptomatic* vs. *vertoonden actieve symptomen*). Other translation shifts, i.e. structural shifts may also cause a different labeling, especially when a scientific term is translated with a relative clause (e.g. *pretreated* vs. *die reeds behandeld zijn*).

- Another reason for this difference in labeling may be provided by the fact that word groups can be considered as multi-word terms in one language, and as separate words or terms in the other language. This problem often relates to the differences in compounding rules between English and Dutch. In Dutch, most compounds are written in one word, whereas English has a tendency towards separating the different components of a compound. This is exemplified by *symptom scores*, which is translated as *symptomenscore*. *Symptom* and *score* are labeled as popular, but their combined use transforms them into a multi-word term.

- Finally, there are some morphologically complex terms in English which have a pure Dutch translation (e.g. *pre-existing* vs. *bestaand*). It is self-evident that these complex terms are more likely to be perceived as scientific terms. This phenomenon is also observed in the opposite direction (e.g. *randomized* vs. *aselect*). Another reason for the mismatched labeling may also be the lack of morphological transparency of some terms. *Renal* (Dutch: *nier-*), for example, is an adjective which means 'pertaining to the kidney (Dutch: *nier*)'. Or *dental* (Dutch: *tand-*) relates to the noun *tooth* (Dutch: *tand*). In these cases, the morphological relation between noun and adjective is not obvious. These terms are also more likely to be perceived as scientific (e.g. *renal disorders* vs. *nieraandoeningen*).

We will now continue with a description of the term extraction experiments, which aim for an accurate detection of scientific terms in the EPAR data sets.

# 3 An Endo- and Exogenous Baseline

In order to assess the complexity of medical term detection in patient leaflets, we experimented with two baseline approaches, of which the first relies completely on external information sources, whereas the latter solely relies on text-internal features surrounding the word of interest.

## 3.1 Lexicon-based Term Extraction

The most straightforward procedure for the detection of scientific terms in the EPAR corpus consists of a dictionary-based or lexicon-based lookup: each sequence of words in the text that matches an entry in the lexical resources is considered as a term occurrence. The following medical lexical resources were used for the term extraction:

- `MeSH`: The Medical Subject Headings thesaurus is a controlled vocabulary, produced by the National Library of Medicine, and used for indexing, and searching for biomedical and health-related information and documents in English. The thesaurus is available from http://www.nlm.nih.gov/mesh in various formats. MeSH consists of 22,995 descriptors, which are organized in 15 hierarchies and a set of 150,000 supplementary concept records. Terms can occur in different hierarchies.

  For the experiments, we selected all unique English entries, which resulted in a lexicon of 23,859 entries representing 24,280 unique terms (an entry can contain multiple terms). For the Dutch version of the MeSH, we relied on a termbase described by [4]. This translation project focuses mainly on chapters C and E (Diseases and Analytical, Diagnostic and Therapeutic Techniques and Equipment respectively) and contains 4,355 unique entries. Taking into account the synonyms which are given for some IDs, a total lexicon size of 5,344 word forms was obtained.

- `Taalvlinder + Ziekenhuis.nl`: As the Dutch MeSH termbase is rather limited in size and scope, the use of extra Dutch lexical resources was imperative. Two very deserving medical lexicons, Taalvlinder[3] and the Ziekenhuis.nl dictionary by Medical Media[4] provided us with 4875 extra Dutch and 4921 extra English items for the detection of scientific terminology.

However, despite being focused on the medical domain, these medical resources provide no information on the (lack of ) scientific character of a given term. Neither MeSH nor Ziekenhuis.nl differentiate between scientific terms such as *12-Hydroxy-5,8,10,14-eicosatetraenoic Acid* or *sebaceous adenocarcinoma* and more general terms such as *wild animals* or *pregnancy.* This implies, for example, that the MeSH heading *headache* [C10.597.617.470], makes no distinction between synonymous scientific terms *cephalalgia*, *cephalgia*, *hemicrania* and the more popular term *head pain.* In order to filter out these popular terms, we used the `Celex`[2] database for Northern Dutch as a filter. Celex is a compilation of the Van Dale's Comprehensive Dictionary of Contemporary Dutch (1984), the Word List of the Dutch Language (1954; revised version) and the most frequent lemmata from the text corpus of the Institute for Dutch Lexicology (INL). It contains frequency information, and phonological, morphological, and syntactic lexical information for more than 380,000 word forms. The English version which is based on the Oxford Advanced Learner's Dictionary (1974) and the Longman Dictionary of Contemporary English (1978) provides information on more than 160,000 word forms. For the experiments, we used the word form files for both languages.

For the automatic recognition of the scientific terms, we concatenated the medical lexicons MeSH, Taalvlinder and Ziekenhuis.nl; for filering out the popular terms from the medical lexicons we took the intersection of the resulting data sets and the Celex lexical database and we kept for further processing all medical terms which did not occur in Celex. Since we consider the popular character of a term also being linked to its frequency, all Celex terms with a frequency of >10 were taken into account for filtering. This implies that the Dutch word *zygoot* (English: *zygote*), which has a zero frequency, was not considered a popular term. For Dutch, this resulted in the omission of words such as *hallucinatie* (Eng.: *hallucinations*), *halsslagader* (Eng.: *Carotid Arteries*), *halswervel* (Eng.: *Cervical Vertebrae*) and *hartinfarct* (Eng.: *heart attack*). The second an third example also illustrate our findings on Latin-based terms in Section 2.2. For English, terms such as *enzymes*, *epilepsy*, *ether*, *ethics*, etc. were filtered from MeSH. Table 3 gives an overview of the number of unique

[3] available at http://www.ochrid.dds.nl/medici.htm

[4] available at http://www.ziekenhuis.nl/index.php ?cat=woordenboek

terms in medical lexicons before and after intersection.

| | Lexicon | before intersect. | after intersect. |
|---|---|---|---|
| English | MeSH | 24,280 | 23,032 |
| | combined | 29,105 | 27,034 |
| Dutch | MeSH | 5,344 | 4,922 |
| | combined | 9,729 | 8,632 |

**Table 3:** *Number of unique terms in MeSH and the joint medical lexicon before and after intersection.*

Both word forms and lemmata of the two EPAR corpora were matched with the lexicons. As shown in Table 4, this resulted in high precision scores, whereas recall was overall below 50%. As expected, the enlargement of the lexicon leads to an increase of recall, slightly at the cost of precision. Furthermore, the small lexicon for Dutch, which is about 1/5 of the English MeSH lexicon and about 1/3 of the combined English lexicon, results in recall and F-1 scores which are only about 15% below those obtained on the English EPAR data.

| EPARs | | Prec. | Rec. | F=1 |
|---|---|---|---|---|
| English | MeSH | 98.56 | 42.59 | 59.48 |
| | combined | 97.73 | 50.31 | 66.43 |
| Dutch | MeSH | 99.76 | 29.30 | 45.30 |
| | combined | 99.21 | 35.14 | 51.90 |

**Table 4:** *Precision, recall and F=1 scores on the scientific terms in the EPAR corpora.*

A manual inspection of the results with the intersected combined lexicons shows that the low recall is mainly due to the fact that some words are just lacking in the medical lexicon (e.g. *perennial*). For English, 69% (Dutch: 77.4%) of the missed medical terms are not covered by the lexicon. The other errors are caused by a too harsh filtering by Celex (Eng: 31% and Dutch: 22.6% of the errors). There were several causes for filtering out medical terms: Scientific words that have evolved towards common vocabulary, such as *obese*, *oestrogen* and *hepatitis* were filtered by Celex. Furthermore, polysemous words with one scientific meaning, such as *agent* were also filtered by Celex because of high frequencies. Finally, some scientific words have frequencies that are just above the threshold such as *dehydration* (13 occurrences) or *concomitant* (17 occurrences) and were filtered away erroneously.

In order to overcome the low coverage (see also [1]) of this type of exogenous lexicon-based disambiguation, we investigated a baseline machine learning based approach to scientific term extraction, which only relies on text-internal information.

## 3.2 Endogenous Learning-based Term Extraction

As a second baseline, we experimented with a machine learning approach which does not include any text-external lexical resources, nor the word (lemma and part-of-speech) of interest itself. The use of machine learning approaches to automatic term extraction has already been explored in for example biomedical term extraction (see for example [5]). In a machine learning approach, training data are used to learn features that are useful and relevant for automatic term recognition and classification. In this paper, we investigate how memory-based learning (MBL) approach can be applied to the automatic detection of scientific versus popular terms in EPARs. 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 (Lazy learning). At classification time, the algorithm classifies new instances by searching for the nearest neighbors to the new instance using a similarity metric, and extrapolating from their class. In our experiments we use the TIMBL [7] software package that implements a version of the *k* nearest neighbour algorithm optimised for working with linguistic datasets and that provides several similarity metrics and variations of the basic algorithm. The choice in favour of MBL can be motivated by the observation in previous work [9] that TIMBL is quite robust in case of a largely imbalanced class distribution, such as the one we are confronted with in this experiment. In a lazy learning approach, all examples are stored in memory and no attempt is made to simplify the model by eliminating low frequency events, which would be harmful in this type of imbalanced data sets.

For the baseline experiment, in which we used the learner in its default settings, we selected the following basic features for disambiguation:

- **Morpho-lexical local context features** which provide information on the word form, lemma and part-of-speech of two words before and after the focus word.
- **Orthographic features** which inform on the presence or absence of numeric symbols in the terms and which inform on the use of multiple capital letters in one word.
- Two **trigram features** which represent the initial and final trigram of a given word.

The learner had to differentiate between three classes: “scientific”, “scientific_ambig” and ”popular”. The scientific category represents the terms on which both annotators agreed; this was also our category of interest in the evaluation in the previous section. The scientific_ambig class, on the other hand, represents the words which received a scientific label by one of the two annotators. The reason for this annotation was double. We wanted to see whether the disagreement of both annotators was reflected by a lower accuracy on this category. Furthermore, reformulating the learning task as a binary classification task would have led to an arbitrary addition of this ambiguous class to the scientific or popular class. Finally, the popular category covers all other, i.e. the ‘medium’ and the ‘popular’ annotations.

For the experiments, we performed *k*-fold cross-validation on the data sets, which implies hat the data is split into $k$ subsets. Iteratively, each portion is used as a hold-out test set, whereas the remaining ($k$ - 1)/$k$ balance of the data is used for training. For our experiments, $k$ was set to 20, the number of documents in each data set. Table 5 gives an overview of the performance of the baseline TIMBL on the three categories. As opposed to the lexicon-based approach, which showed a large performance difference between both languages and which contrasted high precision scores with low recall scores, the learning-based approach reveals similar and more balanced results for both languages. For both Dutch and English, the learning approach which is based on endogenous information and which does not use word, lemma and POS information on the focus word, yields an F-score of about 65% (Eng: 64.7% and Dutch: 66.9%). In both experiments, the two trigram features had the highest informativeness values (gain ratio=0.06 and 0.05).

| | | Prec. | Rec. | F=1 |
|---|---|---|---|---|
| English | | | | |
| | scientific | 69.66 | 60.44 | 64.72 |
| | scientific_ambig | 41.09 | 33.41 | 36.85 |
| | popular | 95.25 | 97.36 | 96.29 |
| Dutch | | | | |
| | scientific | 73.59 | 61.32 | 66.90 |
| | scientific_ambig | 40.21 | 31.98 | 35.63 |
| | popular | 95.54 | 97.81 | 96.67 |

**Table 5:** *Precision, recall and F=1 scores of* TIMBL *on the three classes. The feature vector does not incorporate the features based on external lexical sources nor the word, lemma and POS feature describing the focus word.*

# 4 Optimized Hybrid Term Extraction

Having explored two extreme perspectives in the baselines, we opted for an optimized hybrid learning-based term extraction. In order to overcome the low coverage of the lexicon-based pattern-matching approach, we included our lexical resources as two additional features in the feature vector: one feature which informs on the presence or absence of the word in the language-specific combined medical lexicon (MeSH, Taalvlinder and Ziekenhuis.nl) and a second feature which checks for the presence or absence of the word in the CELEX lexicon. Furthermore, we also included morpho-lexical features, which give information on the focus word itself (word form, lemma and part-of speech information).

## 4.1 Experimental setup

For the experiments, we again performed 20-fold cross-validation on the data sets. Since the outcome of a machine learning experiment can be strongly influenced by for example the data set used, its internal class distribution, the information sources and the parameters of the learner (see for example [3] or [6]), we ran an internal 19-fold cross-validation optimization loop on the training data for joint feature selection and parameter optimization. Although TIMBL provides sensible default settings which are evaluated on a number of NLP tasks, it is by no means certain that they will be the optimal parameter settings for our task of scientific terminology extraction. Furthermore, although the learner incorporates different feature weighting metrics, such as information gain, gain ratio [15] and chi-squared weighting [17], learning speed and classification accuracy can still be negatively influenced by features which add no or little information beyond the information provided by the other features.

In order to manage the computational expense of joint feature selection and parameter optimization, optimization was performed by means of a genetic algorithm (GA). We used a generational GA with a population of 10 individuals over a maximal number of 30 generations, using uniform crossover (rate: 0.9), tournament selection (size: 2) and discrete and Gaussian mutation on the features. For our experiments, each individual is represented as a string and contains particular values for all algorithm parameters (see [7]) and for the selection of the 21 features which are represented in the chromosome as ternary alleles (ignore, weighted overlap or modified value difference metric). Figure 1 gives a schematic view of an example individual. In order to decide which individuals will survive into the next generation, we opted for the F-score on the "scientific" class, which is our main class of interest, as fitness function.

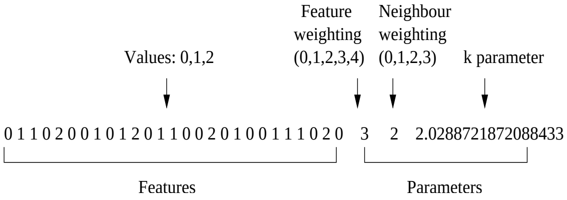

**Fig. 1:** *A GA individual with particular values for the features and the algorithm parameters.*

The internal 19-fold cross-validation loop led to the selection of one optimal individual for each of the 20 test folds. If we consider the predominant feature selection and the informativeness of the selected features after optimization, the following could be observed for Dutch and English. Feature selection has mainly led to filtering out some lemma and POS information of the surrounding words. For Dutch, the gain ratio values of the remaining selected features show that the highest value is assigned to the feature which informs on the presence or absence of the word in the language-specific combined medical lexicon consisting of MeSH, Taalvlinder and Ziekenhuis.nl (information gain=0.1, gain ratio=0.5). Strangely enough, this feature is completely filtered out in the optimal setting for English. Further informative features for both languages are the ones checking for the presence or absence of the word in the CELEX lexicon, followed by the two trigram features and the lemma and POS feature of the focus word.

### 4.2 Results on the Hold-out Test Data

Table 6 gives an overview of the performance of the GA optimized TIMBL on the three categories. It yields highly accurate results on the predominant (> 90% of the instances) popular class. As expected, the lowest results, i.e. an F-score of 41% for Dutch and 45% for English, are obtained for the ambiguous, and low frequent (ca. 3% of the instances) scientific_ambig class. The results on the scientific class show that a learning-based approach compares favorably to the lexicon-based approach. Both precision and recall scores are more than 20% higher, whereas the lexicon-based approach suffered from low recall scores. Furthermore, Table 6 shows similar results for both languages, in contrast to the lexicon-based approach which was not stable across the two languages.

| | | Prec. | Rec. | F=1 |
|---|---|---|---|---|
| English | | | | |
| | scientific | 80.28 | 75.78 | 77.96 |
| | scientific_ambig | 61.01 | 36.17 | 45.42 |
| | popular | 96.24 | 99.01 | 97.61 |
| Dutch | | | | |
| | scientific | 83.45 | 77.29 | 80.25 |
| | scientific_ambig | 58.81 | 31.64 | 41.14 |
| | popular | 96.95 | 99.38 | 98.15 |

**Table 6:** *Optimized precision, recall and F=1 scores of* TIMBL *on the three classes.*

In order to discover regularities in the errors committed by TIMBL, we also performed a manual error analysis on the "scientific" class in both languages.

## 5 Qualitative error analysis

For the manual error analysis, we started from the confusion matrices in Tables 7 and 8, which show similar tendencies.

| | scient. | scient.<br>ambig | pop. | total |
|---|---|---|---|---|
| English | | | | |
| scientific | **973** | 194 | 45 | 1212 |
| scient._ambig | 102 | 327 | 107 | 536 |
| popular | 209 | 383 | 15,171 | 15,763 |
| total | 1284 | 904 | 15,323 | 17,511 |
| Dutch | | | | |
| scientific | **1099** | 178 | 39 | 1316 |
| scient._ambig | 78 | 187 | 53 | 318 |
| popular | 245 | 226 | 14,988 | 15,459 |
| total | 1422 | 591 | 15,080 | 17,093 |

**Table 7:** *Confusion matrix showing the number of words per error class for Dutch and English. Column labels are referring to manual annotation whereas row labels refer to system output.*

Looking into more detail into the different error classes, a number of observations can be made.

- **Popular terms being predicted as scientific or scientific ambiguous:** The items belonging to these classes (1% of all words and 14% of all wrong labels in English) share a number of characteristics. The majority are morphologically more complex words, such as *post-autorisation*, *animal-derived*, *short-acting*, *anti-sickness* and less frequent words (e.g *deficiency*, *cartridges*). Another category are those words that annotators haven't labeled as scientific because they can't be replaced by a popular variant, such as acronyms (e.g *RH12*, *M05*) or proper names referring to pharmaceutical companies (e.g *GlaxoSmithKline*). A possible solution for this problem would be a postprocessing step that removes scientific labels for words occurring in a specialised EPAR lexicon containing frequently used proper names and acronyms. Another subclass points to real labeling errors, such as *biochemical*, *chemotherapy*, *diagnosis*, *receptor*, *hypersensitivity*). Only a small set of words reveals real prediction errors, such as *blood*, *place*, *active*, *type* and *II*. A number of these words form a scientific term together with the preceding or following word, which might explain why annotators have sometimes labeled these as scientific (e.g *type II*).

| | scient. | scient.<br>ambig | pop. | total |
|---|---|---|---|---|
| English | | | | |
| scientific | **275** | 92 | 42 | 409 |
| scient._ambig | 55 | 73 | 64 | 192 |
| popular | 146 | 235 | 1459 | 1840 |
| total | 476 | 400 | 1565 | 2441 |
| Dutch | | | | |
| scientific | **326** | 105 | 38 | 469 |
| scient._ambig | 47 | 44 | 43 | 134 |
| popular | 186 | 156 | 1703 | 2045 |
| total | 559 | 305 | 1784 | 2648 |

**Table 8:** *Confusion matrix showing the number of types (unique words) per error class for Dutch and English.*

- **Mismatch between scientific and scientific ambiguous:** Labeling a scientific word as scientific ambiguous and vice versa (2% of all words and 28% of all wrong labels in English) is probably not as damaging as labeling it as popular. This hypothesis is confirmed when inspecting the items of both these classes. Most of the items are real scientific terms such as *pharmacodynamics*, *diabetic*, *glucose*, *insulin*, *hypersensitive*, *schizophrenia*, an observation which would probably justify a merger of the scientific and scientific ambiguous classes.

- **Scientific and scientific ambiguous terms being predicted as popular:** This error class is the most problematic one, both quantitatively (3.5% of all words and 57% of all wrong labels in English) and qualitatively (given the final goal of this study that consists in replacing scientific terms by their popular variant). A large number of the items belonging to this class are real scientific terms where our approach fails (e.g. *oncology*, *dehydration*, *episcleritis*, *oestrogen*). There are a number of possible reasons for this failure:

– The skewed class distribution: as the majority “popular” class covers over 87% of the instances in our training set, the learned algorithm is biased towards this majority class label.

– The very limited feature set which does not always allow for disambiguation: manual inspection of items that are both predicted correctly and incorrectly (e.g *coagulation*, *osteoporosis*, *congestion*, *dialysis*) reveals a big feature overlap in both cases. This could be solved by a more exhaustive feature set which incorporates more linguistic (e.g. chunk) information, morphological information, a full trigram composition, etc. , together with a larger and more varied training corpus for achieving better results.

– The limited lexicon: the lexicon feature also provoques a number of problems that have already been described in Section 3, which could partially be solved by the inclusion of more lexical resources, such as for example MedDRA (Medical Dictionary for Regulatory Activities). The current low added value of this lexical information might explain that this feature was even filtered out in the final English learning experiment.

– The scope of the scientific terms: in our approach we have mainly focussed on isolated words, assuming that multiword terms would be retrieved implicitely by considering the local context as an important feature. The performance on 2-word terms (e.g *insulinedependent diabetes*, *protease inhibitor*) seems reasonable as we haven’t done anything special for these multiword terms (32% labeled correctly on a total of 115). If we also take into account the scientific ambiguous labels, the improvement is considerable (54% labeled correctly). The figures for 3- and 4-word terms such as *body mass index* look much worse. Here, we only retrieve 4 correct instances on a total of 27 terms. This could be solved by incorporating chunk information in the feature vector and by taking into account previous decisions into the decision process for a given instance.

## 6 Concluding Remarks

In this paper, we investigated the presence of scientific terms in a patient information corpus of EPARs. Annotation experiments on a parallel corpus of Dutch and English terms showed that about 10% of the words in both languages can be labeled as scientific. In order to assess the complexity of medical term detection in patient leaflets, we experimented with two baseline approaches and an optimized hybrid learning approach. We showed that the learning approach which relied on a limited number of straightforward features and which did not use any lexical information outperformed the lexicon-based approach. The hybrid learner obtained an F-score of about 80% for both languages. However, the majority of the errors committed by the learner involves a false ‘popular’ classification for a scientific term, which might be due to the small and imbalanced data set, the rather rudimentary feature set and the low coverage of the existing medical lexicons. We plan to further investigate each of these sources of errors. In a next step, we plan the automatic replacement of these scientific terms by their popular alternative and the evaluation of the (improved) readability of the resulting patient information by a balanced patient group.

## References

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

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

[3] M. Banko and E. Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-01)*, pages 26–33, 2001.

[4] J. Buysschaert. 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*, pages 39–43, 2006.

[5] N. Collier, C. Nobata, and J. Tsujii. Extracting the names of genes and gene products with a hidden markov model. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 201–207, 2000.

[6] W. Daelemans, V. Hoste, F. De Meulder, and B. Naudts. 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)*, pages 84–95, 2003.

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

[8] I. Dagan and K. Church. Termight: identifying and translating technical terminology. In *Proceedings of Applied Language Processing*, pages 34–40, 1994.

[9] V. Hoste. *Optimization Issues in Machine Learning of Coreference Resolution.* PhD thesis, Antwerp University, 2005.

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

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

[12] A. Maynard and S. Ananiadou. Identifying terms by their family and friends. In *Proceedings of COLING-2000*, pages 530–536, 2000.

[13] K. Nink and H. Schroder. *Zu Risiken und Nebenwirkungen/lesen Sie die Packungsbeilage?* Bonn: Wissenschaftliches Institut der AOK, 2005.

[14] P. Pantel and D. Lin. 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, 2001.

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

[16] L. Van Vaerenbergh. 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, 2007.

[17] A. White and W. Liu. Bias in information-based measures in decision tree induction. *Machine Learning*, 15(3):321–329, 1994.

[18] K. Zethsen. Latin-based terms. true or false friends? *Target*, 16(1):125–142, 2004.