ABOP, Automatic Optimization of Patient Information Leaflets

Isabelle Delaere*,§, Véronique Hoste*,§, Claudia Peersman†, Leona Van Vaerenbergh† and Peter Velaerts*,§
*LT^3, Language and Translation Technology Team, University College Ghent
isabelle.delaere@hogent.be; veronique.hoste@hogent.be; peter.velaerts@hogent.be
§Dpt. of Applied Mathematics and Computer Science, Ghent University
†Artesis Hogeschool Antwerpen
claudia.peersman@artesis.be; leona.vanvaerenbergh@artesis.be

Abstract

ABOP, an Automatic optimizer for Patient Information Leaflets (PILs), aims to improve the readability of Dutch PILs by tackling three of the issues that make a PIL hard to read: the scientific terminology used, the redundancy which makes a PIL needlessly lengthy and the overlap between illocutionary acts which often make no distinction between an instruction, a warning and mere informative text. ABOP combines a highly accurate learning-based terminology extraction web service with an application that can be plugged into Microsoft Word and easily used by PIL authors.

1. Introduction

There is a clear need for comprehensible language in patient information leaflets (PILs). Legislative initiatives such as the European Directive (2001/83/EC) on the readability of PILs have been elaborated in order to improve their clarity. Research ([1], [2]) has also shown that patients often feel distressed by reading the PIL which, in some cases, even leads to unsafe behavior. In this paper, we will discuss the first component of the ABOP project, i.e. our approach to automatically detect scientific terminology and to replace it by popular language. Previously, this terminology has proven to be one of the main causes of distress when patients read a PIL [2]. Despite the European Directive mentioned above, current patient information still contains a large number of scientific terms. These terms are often copied from the so-called SPC (Summary of Product Characteristics), a document written by experts in very specific, technical language.

Given this problem's social and economic relevance, we aim to develop an Automatic optimizer of Patient Information Leaflets (ABOP). It is an authoring environment which guides the author through the creation of new leaflet texts as well as the adaptation of existing leaflets. The main objective is not only to produce more consistent and comprehensible patient information but also to do so in a less time consuming way. Apart from guiding the author through the different formal aspect possibilities, ABOP should also give support regarding content and three main problem areas:

- Scientific terminology will be reduced to a minimum and if a popular counterpart exists, it will be given preference over the scientific term. If no popular term is available, the scientific term will be explained by means of a definition or a description
- Redundancy will be eliminated
- The different illocutionary forms will stand out clearly (e.g. through layout) and phrasings will be used applying a "standard expression method", thus leaving no doubt about the risk level

In this paper, we will elaborate on our approach to remove the scientific terminology in PILs and to replace it with popular variants or definitions/descriptions. By doing so, we wish to improve the PIL's readability. To adequately select scientific terminology which requires replacement by a popular variant, we experimented with a machine learning-based approach trained on an annotated corpus. For our experiments, we used an annotated corpus of 625 PILs from different categories of medicinal areas. Our results show a weighted average F-score of nearly 80% for the detection of those terms that require replacement.

The remainder of this paper is structured as follows: Section 2 gives an overview of related research and a motivation for this particular study. In Section 3 we will present our corpus and discuss our annotation guidelines. In Section 4 we will discuss our approach to this problem, i.e. the learning-based term extraction and we will give a detailed overview of the experiments and the feature vector construction, the results of which are discussed in Section 5. In Section 6 we show the practical implementation of this research. Finally, Section 7 concludes this paper and gives some ideas for future work.

2. Background and related work

Much research has been carried out to examine to what extent patients actually understand the information that

is being provided to them. Research in 21 hospitals in the Netherlands [3] has shown that 5.6 % of all acute hospitalizations (509 people) are medicine-related and potentially avoidable. Furthermore, the report states that errors with regard to dosage, administering and irregular use of the medicine cause more than 30% of these unnecessary hospitalizations. This kind of information should be easily retrievable in the PIL. However, recent research on Dutch patient leaflets [2] has proven that patients often cannot find the information they are looking for. Pander Maat recently conducted research on the readability of three PILs in the Netherlands. He concluded that patients often cannot find the information they are looking for. However, when they do find it, up to 90% of the respondents interpreted the information correctly. This means that understanding the information is less problematic than finding it. Pander Maat further concluded that when the information is not understood, it is often related with the use of scientific terminology, the lack of instructions and dense paragraphs. Sless and Wiseman [4] already described that readability testing on PILs should "test the accessibility, comprehensibility or the capacity of the participants to act appropriately on the information".

3. Corpus

A corpus of 625 PILs was collected for our experiments. These PILs were taken from various categories such as medicines for the cardiovascular system and the nerve system, hormones, medicines to treat infections, etc. Our corpus not only contains PILs from the European Medicines Agency, but also Belgian PILs that were registered in Belgium.

3.1. Annotation

Our main objective is to replace scientific terms by their popular counterparts in order to obtain a PIL that is comprehensible to the average layman as it is described in the EMEA directive[1]. If no popular counterpart is available, our system suggests a definition or a description of the term that requires replacement. In order to be able to automatically detect terms which come into account for replacement, a linguist annotated the corpus according to strict annotation guidelines. We used Callisto (http://callisto.mitre.org) as an annotation environment and we differentiated between 8 main labels, viz.:

- **scient(ific)**: scientific expert terms such as *neuralgie (neuralgia)*, *angina pectoris* and *rhinitis* which need a replacement or a description
- **scient(ific)_abbr(eviation)**: scientific abbreviations such as *NSAIDs* that need a replacement or a description
- **scient(ific)_pop(ular)**: scientific terms for which a lesser known popular variant exists. (e.g. *anorexia* vs. *magerzucht*)
- **pop(ular)_scient(ific)**: popular variants whose scientific counterpart is better known (e.g. *overgevoelig* vs. *allergisch*)
- **amb(iguous)**: terms with a scientific nature that are widely known to the general public. (e.g. *antibiotics*)
- **amb(iguous)_abbr(eviation)**: abbreviations with a scientific nature that are widely known to the general public. (e.g. *HIV*)
- **namedEntity**: (e.g. *Aspirin*)
- **sub(stance)**: scientific terms to indicate substances such as *silicon dioxide*

The idea behind the different categories is the following; we aim to replace every scientific term by its popular counterpart wherever possible. If no such popular counterpart is available (e.g. in the case of an abbreviation), we aim to add a description or a definition. The categories containing terms that require a replacement or an addition are: *scientific, scientific abbreviations* and *popular scientific*. Terms from this last class, i.e the *popular scientific* terms, are also replaced because the scientific term is actually better known than the popular variant in this particular case or because the popular variant does not completely cover the content of the scientific term (e.g. *depressie* vs. *zwaarmoedigheid*). Terms from the other classes do not require replacement because they are either widely known (viz.: *scientific popular*, *ambiguous*, *ambiguous abbreviations*), cannot be replaced (*substances*) or need not be replaced (*namedEntities*). This fine-grained classification scheme, allowed us to differentiate between similar categories such as for example *scientific* and *substances*.

In order to measure interannotator agreement, 15 PILs (22,754 tokens) were annotated by four annotators. The resulting kappa-score [5]. was 0.80.

4. Learning-based term detection

In previous research, we experimented with a lexicon-based approach [6] . However, given the low coverage of this dictionary-based approach, we experimented with learning-based term detection which not only takes into account lexical and morphological information, but also local context, frequency, etc.

4.1. Memory-based learning

We experimented with a memory-based learning approach which is based on the memory-based reasoning [7] and case-based reasoning schemes [8], [9] which state that perfor-

1. http://ec.europa.eu/enterprise/pharmaceuticals/eudralex/vol-1/dir_2004_27/dir_2004_27_en.pdf

mance 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, an unknown 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" [10] 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). In our experiments we use the TIMBL [11] software package that implements a version of the k-nn algorithm optimised for working with linguistic datasets and that provides several similarity metrics and variations of the basic algorithm.

4.2. Feature construction

We constructed a rich feature vector set for our experiments.

- **Local Context**
  The local context can provide us with strong indications of the scientific character of a given word. Therefore we added word form, lemma and part of speech (POS) tag information for three words before and three words after our focus word. For this task, we used Tadpole, a modular system integrating a memory-based tagger, lemmatizer and morphological segmenter [12]

- **Lexical information**
  In order to build a lexicon of medical terminology, a large number of sources was needed as there are few existing lexicons for Dutch. We used the Dutch version of the MeSH as described by [13] and Dutch lexicons such as Taalvlinder, Elseviers Medische Encyclopedie, the Wikipedia page "Gezondheid van A tot Z", the Dutch entries in the Multilingual Glossary of technical and popular medical terms[2] and the Dutch entries from the Medical Dictionary for Regulatory Activities, MedDRA[3]. In addition to these lexicons, we used sources such as:
  - **Patients' associations** e.g. CMP Vlaanderen & Dystrofie
  - **Online medical dictionaries** e.g. Maranje
  - **Specific websites** e.g. Dokterdokter.nl

  This resulted in a lexicon which was then intersected with Celex in order to filter out everyday language and to only maintain specific terminology. We used three approaches to use this lexicon (75,000 terms) in our feature vector:
  - **Single Word Terms:** We matched every single word with the lexicon; e.g. *rhinitis*
  - **Multiword Terms:** We matched every n-gram (up to five) with our lexicon; e.g. *diabetes mellitus*
  - **Fuzzy Word Terms:** We matched every entry in our lexicon combined with another entry; e.g. *renal* and *disease* occur in our lexicon so *renal disease* becomes a fuzzy match

  This lexical information was converted into three binary features.

- **Substances**
  In our corpus, we used a tag to annotate substances such as *cetostearylalcohol* as these are terms with a scientific nature which cannot be replaced by a popular counterpart because such a counterpart usually does not exist. Additionally, we created a lexicon of 127,000 unique substances to match our PILs with through a binary feature.

- **TF-IDF**
  We calculated the average TF-IDF values (Term Frequency - Inverse Document Frequency) for every word in our PIL corpus based on a substantial part of the Twente Nieuws Corpus, i.e. nearly 100 million words. This TF-IDF approach is based on two ideas: specific terms have a high frequency in the given document (TF) and a term is more distinctive when it occurs in few documents (IDF). So, given a document collection D, a word w, and an individual document $d \in D$,

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

  where $f_{w,d}$ equals the number of times w appears in d, $|D|$ is the size of the corpus and $f_{w,D}$ equals the number of documents in which w appears in D [14].

- **Cognates**
  During previous research on the EPARs (European Public Assessment Reports)[4], we constructed a list of cognates for English and Dutch. This list was constructed by a manually sentence-aligned English-Dutch EPAR corpus which was then automatically aligned at the word level using the Perl implementation for IBM Model One that is part of the Microsoft Bilingual Sentence Aligner [15]. 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 [16] was calculated for each word pair; it involves finding the longest subsequence common

2. http://users.ugent.be/ rvdstich/eugloss/welcome.html
3. http://www.meddramsso.com/MSSOWeb/index.htm
4. http://www.emea.europa.eu/htms/human/epar/eparintro.htm

to the pair of sequences. We used this list of Dutch cognates to construct a binary feature.

- **Affix information**
  It is well known that medical terminology is greatly influenced by Latin and Greek. In Dutch, more than in English, the comprehensibility of Greco-Latinate forms is rather low and their use leads to terminology that is hard to understand for the average layman. Therefore, we constructed a list of prefixes, suffixes and affixes and used this list to build three binary features to detect those terms that contain a Greco-Latin affix.

- **Orthographic information**
  Within a given word, there may be indicators of its scientific character. Two of these indicators are used in our feature vector as binary features. If a word contains numeric symbols, the probability of it being a scientific term rises e.g. *BRCA1-gen*. Second, a word is often a term (i.e. an abbreviation or an acronym) if it contains multiple capital letters e.g. *PET-scan*. The initial and final trigram may also give an indication of whether a given word is scientific or not. We included these two trigrams as two features in our vector e.g. *bronchitis: bro & tis*.

- **Indicative patterns**
  Our research on the EPARs showed that the local context of a scientific term such as the name of a disease often contains a certain pattern. Some examples of these patterns are “is aangewezen bij” (is indicated for), “de behandeling van patiënten met” (the treatment of patients with), etc. To use this information, we built a binary feature to indicate whether a given word or multiword was preceded by one of our patterns or not.

5. Experimental results

The features described in Section 4.2 were combined in a feature vector which was fed to the TiMBL memory-based classifier. The classifier was used with its default parameter settings. All experiments were conducted in a 10-fold cross-validation set-up and for all output classes precision (P), recall (R) and $F_{\beta=1}$ were measured.

$$P = \frac{No.\ of\ correctly\ extracted\ terms\ by\ the\ system}{Total\ No.\ of\ extracted\ terms\ by\ the\ system} \quad (2)$$

$$R = \frac{No.\ of\ correctly\ extracted\ terms\ by\ the\ system}{Total\ No.\ of\ actual\ terms\ in\ the\ text} \quad (3)$$

$$F - score = \frac{2(Precision\ *\ Recall)}{Precision\ +\ Recall} \quad (4)$$

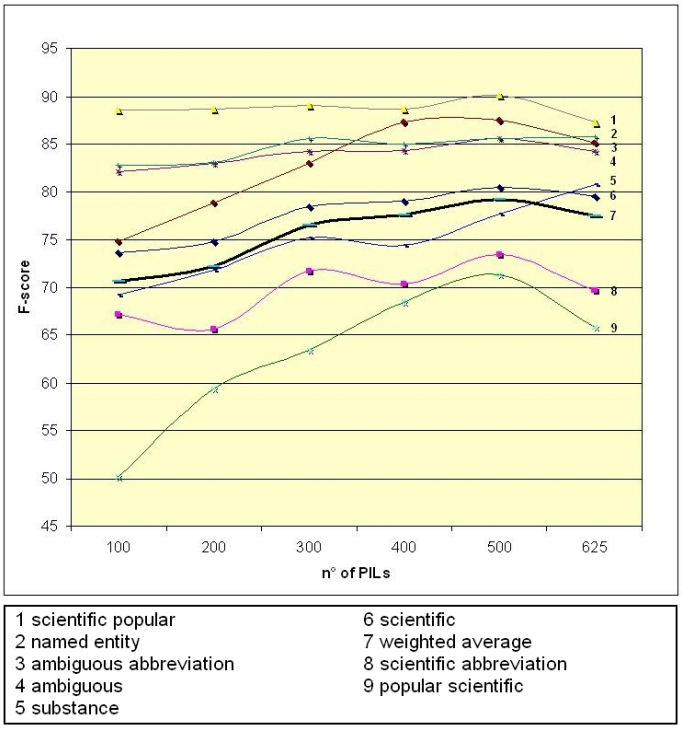

Figure 1. Learning curves

In a first set of experiments, we determined the optimal number of PILs for the classifier by running ten-fold cross-validation experiments with 100, 200, 300, 400, 500 and 625 PILs. Figure 1 gives an overview of the results of these learning curve experiments. Lines 1 to 6 and 8 to 9 show the results per output class, whereas line 7 focuses on the three classes that are candidates for replacement in the ABOP tool. These last results were calculated according to the number of entries per class and thus show a weighted average of these three classes (scientific, scientific_abbreviation and popular_scientific). Except for two classes (substances and namedEntities), there is a clear tendency throughout the evolution of the results. They improve proportionally to the number of PILs and reach a peak at 500 PILs, dropping again at 625 PILs. These results allowed us to set the optimal number of PILs at 500 to train our learning system.

Table 1 gives a more detailed overview of the precision, recall and $F_{\beta=1}$ for all classes on the 500 PIL corpus and it also shows the the weighted average for the three classes that require replacement as explained above. We obtain an F-score of almost 80% for this average.

A thorough manual error analysis of all classes resulted in

a number of conclusions:

- Due to the large corpus, annotation was not always consistent, which inevitably leads to a deterioration of the results
- Everyday words that are part of a larger, scientific multiword term, were often not determined as scientific, but as popular or ambiguous words, e.g. *allergic* in "allergic rhinitis".
- A number of scientific terms were given the *scientific_popular* label by our system, based on morphological features, e.g. *allergic* (scient_pop) vs. *genetic* (scient)
- Diseases written with a capital letter are sometimes erroneously classified as Named Entities

Class	Precision	Recall	$F_{\beta=1}$
scient(ific)	81.35	77.78	79.52
scient(ific)_abbr(eviation)	70.21	69.07	69.63
scient(ific)_pop(ular)	87.66	86.88	87.27
pop(ular)_scient(ific)	66.85	64.82	65.82
amb(iguous)	83.78	84.56	84.17
amb(iguous)_abbr(eviation)	82.94	87.30	85.06
namedEntity	87.67	84.09	85.84
sub(stance)	80.55	80.90	80.72
weighted average	80.75	77.83	79.25

Table 1. 10-fold cross-validation results on the 500 PIL corpus

6. Practical implementation

The result of this research is an application that can be plugged into Microsoft Word, screenshots of which can be seen in Figures 2 to 5. In the examples, the scientific term *glaucoom* (*glaucoma*) is replaced by a popular variant *verhoogde oogboldruk* (*increased eyeball pressure*) The ABOP tool offers solutions for the three problem areas mentioned before, viz. scientific terminology, redundancy and illocutionary acts. This paper only covers the ABOP solution for the first module, i.e. the replacement of scientific terminology.

The intended work flow is as follows. The author can load a PIL that needs rewriting into Microsoft Word; via a web service, the document is scanned for medical terminology. All terms, which are classified as belonging to one of the three previously mentioned categories which need replacement, are highlighted (screenshot 1) and for these terms a solution is offered. Through the use of contextual icons within the document, ABOP has two ways of offering a solution to the PIL author:

- Replacement by or addition of a popular counterpart
- Addition of a definition or description

These popular counterparts and definitions were manually gathered from a wide number of sources. Evidently, the number of these popular counterparts and definitions is limited to the lexicon we have collected: over 5200 popular variants and over 14,000 definitions or descriptions. Therefore, if a PIL is loaded into ABOP and it contains a medical term which has no solution in our database, we use the Google SOAP API and obtain up to 200 snippets for this particular term.

On the basis of the annotated PILs, we extracted a set of high precision patterns which show a scientific term in combination with a popular counterpart or a definition. Using these patterns, the snippets are consulted for replacement candidates. Similar results are filtered out and a final score is calculated based on the expected success of a pattern and the frequency of candidates in the snippets. The author is offered a list of solutions obtained through this Google search and chooses the solution he prefers. Should no solution be found, the author can consult the entire collection of snippets or even new web pages.

7. Conclusion

With this research, we aimed to construct a learning system which detects terminology in patient information. Our feature vector, which is not only based on lexical information, but also on morphological, orthographic, statistical information etc., was the basis for a system that obtained an average weighted F-score of 80% for those classes we want to replace by a popular variant or a description.

To validate our system, we will process the results of a thorough readability test with a large test group we recently carried out. Furthermore, we will validate our external lexicons of medical terminology to guarantee their contents.

Acknowledgment

This research is carried out within the framework of the IWT-TETRA funded ABOP (Automatic optimizer of Patient Information Leaflets) project (IWT-code: 70103).

References

[1] K. Nink and H. Schroder, "Zu risiken und nebenwirkungen: Lesen sie die packungsbeilage?" Wissenschaftliches Institut der AOK (WIdO), WIdO-Materialien Bd. 53, Tech. Rep., 2005.

[2] H. Pander Maat. (2008) Hoe (on)leesbaar zijn geneesmiddelenbijsluiters? een test van drie veel gebruikte bijsluiters. [Online]. Available: http://www.let.uu.nl/~Henk.PanderMaat/personal/begrijpelijkheid

[3] P. Van den Bemt, T. Egberts, and A. Leendertse. Hospital admissions related to medication (harm). een prospectief, multicenter onderzoek naar geneesmiddel gerelateerde ziekenhuisopnames. [Online]. Available: http://www.knmp.nl/download-bestanden/knmp-vandaag-2/publicaties-knmp/nieuws-en.../eindrapport_harm-nov-2006.pdf

[4] D. Sless and R. Wiseman, *Writing about medicines for people. Usability Guidelines for Consumer Medicine Information.* Commonwealth of Australia, 1997.

[5] J. C. Carletta, "Assessing agreement on classification tasks: the kappa statistic." *Computational Linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[6] V. Hoste, E. Lefever, K. Vanopstal, and I. Delaere, "Learning-based detection of scientific terms in patient information," in *Proceedings of the sixth international conference on Language Resources and Evaluation*, 2008.

[7] C. Stanfill and D. Waltz, "Toward memory-based reasoning," *Communications of the ACM*, vol. 29, no. 12, pp. 1213–1228, 1986.

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

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

[10] D. Aha, "Lazy learning: Special issue editorial," *Artificial Intelligence Review*, vol. 11, pp. 7–11, 1997.

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

[12] A. van den Bosch, B. Busser, S. Canisius, and W. Daelemans, "An efficient memory-based morpho-syntactic tagger and parser for dutch," in *Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting*, 2006, pp. 99–114.

[13] 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*, 2006, pp. 39–43.

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

[15] R. C. Moore, "Fast and accurate sentence alignment of bilingual corpora," in *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, 2002, pp. 135–144.

[16] D. S. Hirschberg, "Algorithms for the longest common subsequence problem," *J. ACM*, vol. 24, no. 4, 1977.

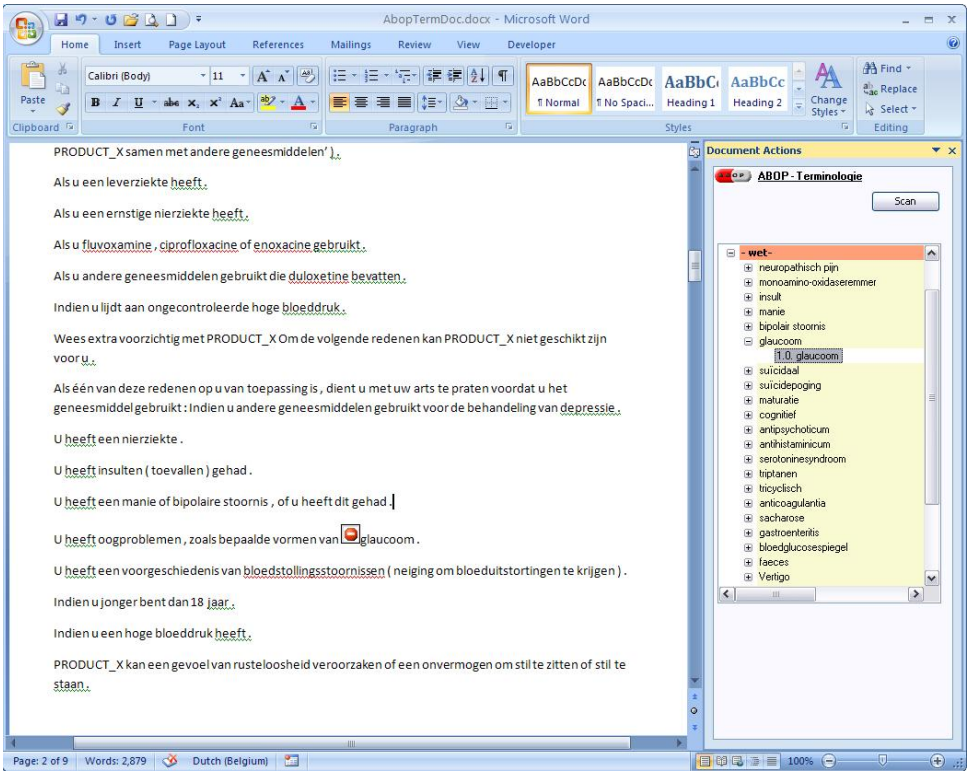

Figure 2. Screenshot 1

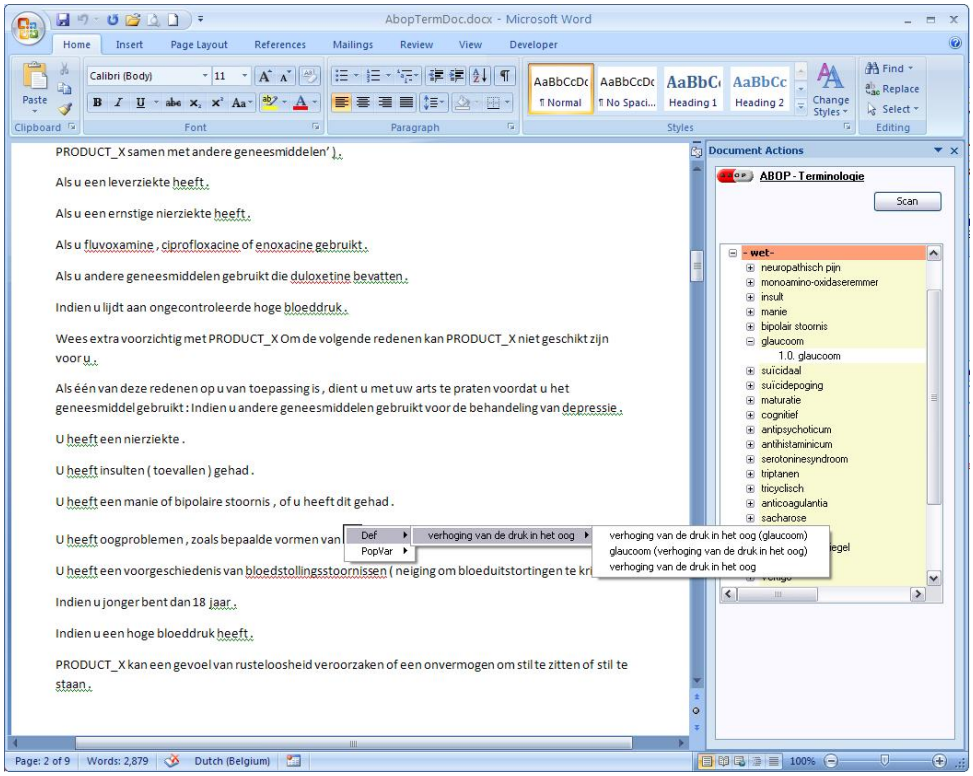

Figure 3. Screenshot 2

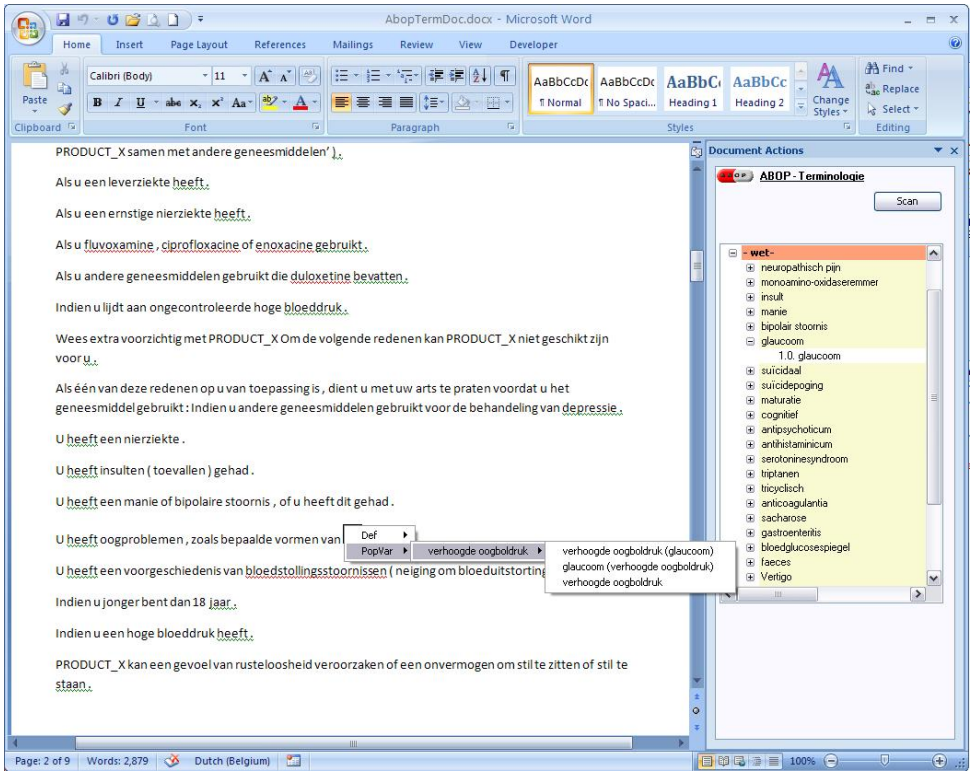

Figure 4. Screenshot 3

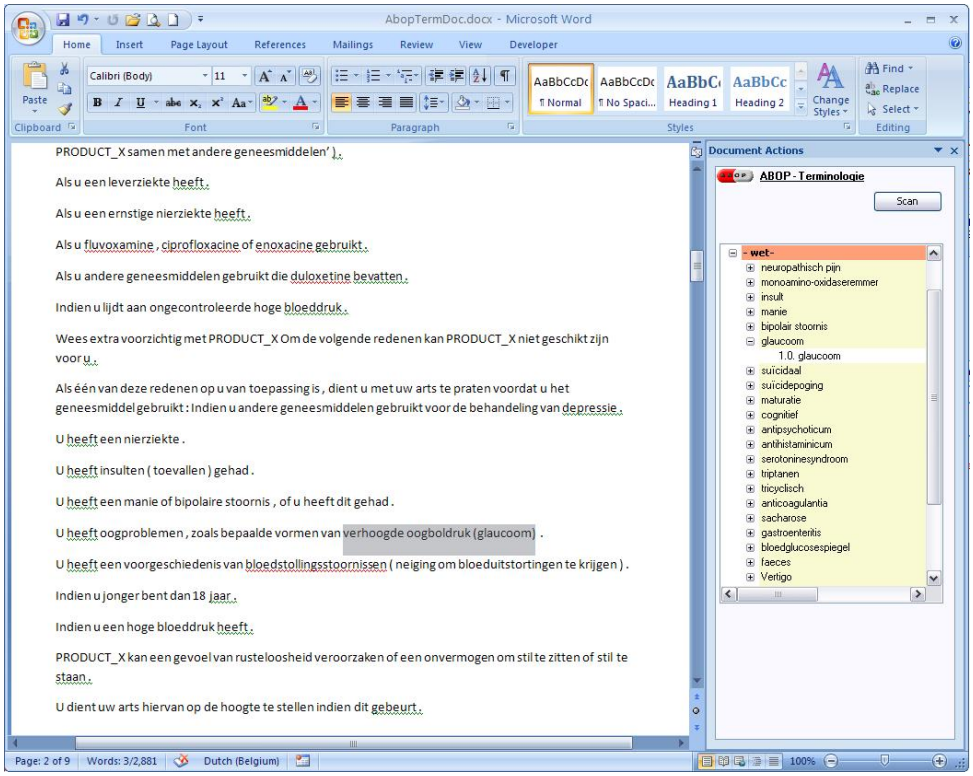

Figure 5. Screenshot 4