UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Detection of Difficult for Understanding Medical Words using Deep Learning

Author:
John SMITH

Supervisor:
Dr. James SMITH

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2018

Declaration of Authorship

I, John SMITH, declare that this thesis titled, "Detection of Difficult for Understanding Medical Words using Deep Learning" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Abstract

Faculty of Applied Sciences

Master of Science

Detection of Difficult for Understanding Medical Words using Deep Learning

by John SMITH

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor...

Contents

List of Figures

List of Tables

List of Abbreviations

NLP	**N**atural **L**anguage **P**rocessing
RNN	**R**ecursive **N**eural **N**etwork
DT	**D**ecision **T**ree **CV**
Cross-**v**alidation	

List of Symbols

a	distance	m
P	power	W ($\mathrm{J\,s^{-1}}$)
ω	angular frequency	rad

For/Dedicated to/To my…

Chapter 1

Introduction

1.1 Motivation

Specialized areas, such as medical area, convey and use technical words, or terms, which are typically related to knowledge developed within these areas. In the medical area, this specific knowledge often corresponds to fundamental medical notions related to disorders, procedures, treatments, human anatomy, etc. For instance, technical terms like *blepharospasm* (abnormal contraction or twitch of the eyelid), *alexithymia* (inability to identify and describe emotions in the self), *appendicectomy* (surgical removal of the vermiform appendix from intestine), or *lombalgia* (low back pain) are frequently used in the medical area texts.

As in any specialized areas, two main kinds of users exist in the medical area:

- medical doctors, both researchers of practitioners, are experts of the domain. They contribute to the creation and development of biomedical knowledge and its exploitation for the healthcare process of patients;
- patients and their relatives are consumers of the healthcare process. Usually, they do not have expert knowledge, while it is important that they understand the purpose and issues of their healthcare process.

If the understanding of technical medical terms is easy for the medical staff, patients and their relatives may present some difficulties in the understanding and using of such terms: they show indeed poor *health literacy*.

Hence, the existing literature provides several studies dedicated to the understanding of medical notions and terms by non-expert users, and on their impact on a successful healthcare process (Mcgray, 2005; Eysenbach, 2007). Yet, it is not uncommon that patients and their relatives must face very technical health documents and information. Examples of this kind are frequent and usually the non-expert users are at loss in such situations:

- understanding of information on drug intake (Vander Stichele, 1999; Patel, Branch, and Arocha, 2002), such as instructions related to the description and specification of steps necessary for the preparation and intake of drugs,
- understanding of clinical documents (Zeng-Treiler et al., 2007), which contain important information on the healthcare process of patients,
- understanding of clinical brochures or informed consents (Williams et al., 1995), which are specifically created for patients and which are typically read by patients during their clinical pathway,

- more generally, understanding of information provided for patients by different websites (Oregon Practice Center, 2008; Brigo et al., 2015) in different languages (English, Spanish, French) and different medical specialties,
- for the same reasons, communication between patients and medical staff (Jucks and Bromme, 2007; Tran et al., 2009) remains complicated.

These various observations provide the main motivation to our work. We propose to address the needs of non-specialized users in the medical domain. As we noticed, the main need is related to the understanding of medical and health information.

1.2 Thesis structure

We first present some related work in chapter 2 and background knowledge in chapter 3 which form the basis of the methods described in this work. We then introduce the data we used in chapter 4 and the proposed method in chapter 5. Our results and their discussion are presented in Chapter 6. Finally, we conclude with some directions for future work in Chapter 7.

Chapter 2

Related Work

Related work is globally related to the detection of technical contents in documents and to their adaptation. Here, we are interested in the first aspect: detection and diagnosis of technical medical contents.

In the NLP (Natural Language Processing) area, work related to the diagnosis of technical medical documents is quite frequent. Traditionally, researchers exploit the readability measures. Among these measures, it is possible to distinguish classical readability measures and computational readability measures (François and Fairon, 2013). Classical measures usually rely on number of letters and/or of syllables a word contains and on linear regression models (Flesch, 1948; Gunning, 1973), while computational readability measures may involve vector models and a great variability of features, among which the following have been used for processing the biomedical documents: combination of classical readability formulas with medical terminologies (Kokkinakis and Toporowska Gronostaj, 2006); n-grams of characters (Poprat, Markó, and Hahn, 2006), stylistic (Grabar, Krivine, and Jaulent, 2007) or discursive (Goeuriot, Grabar, and Daille, 2008) features, lexicon (Miller et al., 2007), morphological features (Chmielik and Grabar, 2011), combinations of different features (Zeng-Treiler et al., 2007).

At a more fine-grained level, the readability of words has been addressed much less frequently. In the general language, some research actions are often performed as part of the NLP challenges, such as the SemEval NLP challenge[1] held in 2012. This challenge proposed the following task: for a short text and a target word, several possible substitutions satisfying the context have also been proposed. The objective was to rate and to order the substitutions according to their degree of simplicity (Specia, Jauhar, and Mihalcea, 2012). The participants applied rule-based and/or machine learning systems. Combinations of various features, designed to detect the simplicity of words, have been used, such as: lexicon from spoken corpus and from Wikipedia, Google n-grams, WordNet (Sinha, 2012); word length, number of syllables, latent semantic analysis, mutual information and word frequency(Jauhar and Specia, 2012); Wikipedia frequency, word length, n-grams of characters and of words, random indexing and syntactic complexity of documents (Johannsen et al., 2012); n-grams and frequency from Wikipedia, Google n-grams (Ligozat et al., 2012); WordNet and word frequency (Amoia and Romanelli, 2012). The best systems reached up to 0.60 Top-rank and 0.575 Recall. Another work has been done on scholar texts in French written for children with the purpose to differentiate between the texts from various scholar levels and to test various features suitable for that (Gala, François, and Fairon, 2013). This system reached up to 0.62 classification accuracy.

In the medical area, we can mention three experiments: manual rating of medical words (Zheng, Milios, and Watters, 2002), automatic rating of medical words on the

[1] http://www.cs.york.ac.uk/semeval-2012

basis of their presence in different vocabularies (Borst et al., 2008), and exploitation of machine learning approach with various features (Grabar, Hamon, and Amiot, 2014). This last experiment achieved up to 0.85 F-measure on individual annotations.

Another issue is to know what are the most suitable data for the analysis of text readability. These data have indeed crutial impact on models created and on their usability. Several approaches have been proposed:

- exploitation of expert judgment, who have an idea on needs of population aimed in the study (Clercq et al., 2014). The main limitation is that experts may have difficulties to figure out what are the real needs of population;
- exploitation of text books created for population according to their readability levels, such as school books (Gala, François, and Fairon, 2013). The main limitation is that such books are usually created by experts using theoretical basis and observations;
- exploitation of crowdsourcing involving large population (Clercq et al., 2014). The main limitation is that the population involved is uncontrolled and unknown;
- exploitation of eye-tracking methods for a more fine-grained analysis of reading difficulties (Yaneva, Temnikova, and Mitkov, 2015; Grabar, Farce, and Sparrow, 2018). The main limitation is that only short text spans can be used;
- manual annotation by human annotators (Grabar and Hamon, 2016). In this case, the annotators represent the population, they are part of the controlled population, they can perform more complicated tasks than in case of crowdsourcing, although they are usually less many than in crowdsourcing experiments.

Related to this issue is the question on generalizability of data and of models generated from these data. For instance, it has been observed that data from experts are difficult to generalize over the population (Clercq et al., 2014).

We propose novel machine learning approaches for a stronger distinction of readability of medical words and distinction of words which may present understanding difficulties to non-experts users. The medical data processed are in French. Seven human annotators participated in creation of the reference data.

Chapter 3

Background Knowledge

- Classification problem, methods to solve: log regression, trees, deep learning
- How to ealuate quality of classification problem (accuracy, precision, recall, F1)
- Word embeddings
- RNNs and how to get embeddings from them

Chapter 4

Dataset description

4.1 Linguistic data description

For the text classification task the data was collected and annotated as described in (Grabar, Hamon, and Amiot, 2014). The source terms are obtained from the medical terminology Snomed International (Côté et al., 1993) in French, available from the ASIP SANTE website[1]. The purpose of this terminology is to provide an extensive description of the medical field. Snomed contains 151,104 medical terms organized into eleven semantic axes such as disorders and abnormalities, procedures, chemical products, living organisms, anatomy, social status, etc. For the purpose of our task, we chose five axes related to the main medical notions: disorders, abnormalities, procedures, functions, and anatomy. Our assumption is that terms in these categories are familiar to a layman, in contrast to contents of such specific groups as chemical products (*hydrogen sulfide*) and living organisms (*Sapromyces, Acholeplasma laidlawii*).

The 104,649 selected terms are lemmatized and tokenized into words (or tokens) resulting in 29,641 unique words such that '*trisulfure d'hydrogène*' provides three words (*trisulfure, de, hydrogène*).

The dataset contains three morphological groups of words:

- compound words which contain several bases: abdominoplastie (abdominoplasty), dermabrasion (dermabrasion);
- constructed words which contain one base and at least one affix: cardiaque (cardiac), acineux (acinic), lipoı?de (lipoid);
- simple words which contain one base, no affixes and possibly infections (when the lemmatization fails): acné (acne), fragment (fragment).

4.2 Annotation process

The set of 29,641 unique words was annotated by seven French speakers, 25-40-year-old, without medical training, without specific medical problems, but with the linguistic background. The annotators are expected to represent the average knowledge of medical words among the population as a whole. The annotators are presented with a list of terms and asked to assign each word to one of the three categories:

- I can understand the word;

[1] http://esante.gouv.fr/services/referentiels/referentiels-d-interoperabilite/snomed-35vf

- I am not sure about the meaning of the word;
- I cannot understand the word.

The assumption is that the words, which are not understandable by the annotators, are also difficult to understand by patients. The annotators were asked not to use dictionaries during the annotation process. The annotation results are represented in Table 4.1 .

Categories	*A1 (%)*	*A2 (%)*	*A3 (%)*	*Un. (%)*	*Maj. (%)*
1. I can understand	8,099 (28)	8,625 (29)	7,529 (25)	5,960 (26)	7,655 (27)
2. I am not sure	1,895 (6)	1,062 (4)	1,431 (5)	61 (0.3)	597 (2)
3. I cannot understand	19,647 (66)	19,954 (67)	20,681 (70)	16,904 (78)	20,511 (71)
Total annotations	29,641	29,641	29,641	22,925	28,763

TABLE 4.1: Number (and percentage) of words assigned to reference categories by three annotators (A1, A2 and A3), and in the derived unanimity (Un.) and majority (Maj.) datasets.

Chapter 5

Methodology

We propose to tackle the described problem through the supervised words categorization. The purpose is to categorize medical words, or terms, according to whether they can be understood or not by non-specialized people. The manual annotations of there words provide the reference data. The categorization pipeline is the following: categorization features are computed, they are used for training the classifiers, and the results are evaluated using the cross-validation. The proposed method has three phases:

1. calculation of NLP features associated with the annotated words;
2. building a machine learning model for words classification;
3. evaluation of classification quality using a cross-validation.

The main research question is whether the NLP methods can distinguish between understandable and nonunderstandable medical words and whether they can diagnose these two categories.

5.1 Feature sets

5.1.1 Standard NLP features

We will refer to previously used NLP features described in (Grabar, Hamon, and Amiot, 2014) as *"standard features"* (opposed to *"embeddings"* described in the next subsection). They include 24 linguistic and extra-linguistic features related to general and specialized languages. The features are computed automatically and can be grouped into ten classes:

- *Syntactic categories.* Syntactic categories and lemmas are computed by TreeTagger (Schmid, 1994) and then checked by Flemm (Namer, 2000). The syntactic categories are assigned to words within the context of their terms. If a given word receives more than one category, the most frequent one is kept as feature. Among the main categories we find for instance nouns, adjectives, proper names, verbs and abbreviations.
- *Presence of words in reference lexica.* We exploit two reference lexica of the French language: TLFi[1] and *lexique.org*[2]. TLFi is a dictionary of the French language covering XIX and XX centuries. It contains almost 100,000 entries. *lexique.org* is a lexicon created for psycholinguistic experiments. It contains over 135,000 entries, among which inflectional forms of verbs, adjectives and nouns. It contains almost 35,000 lemmas.

[1] http://www.atilf.fr/
[2] http://www.lexique.org/

- *Frequency of words through a non specialized search engine.* For each word, we query the Google search engine in order to know its frequency attested on the web.
- *Frequency of words in the medical terminology.* We also compute the frequency of words in the medical terminology Snomed International.
- *Number and types of semantic categories associated to words.* We exploit the information on the semantic categories of Snomed International.
- *Length of words in number of their characters and syllables.* For each word, we compute the number of its characters and syllables.
- *Number of bases and affixes.* Each lemma is analyzed by the morphological analyzer Dérif (Namer and Zweigenbaum, 2004), adapted to the treatment of medical words. It performs the decomposition of lemmas into bases and affixes known in its database and it provides also semantic explanation of the analyzed lexemes. We exploit the morphological decomposition information (number of affixes and bases).
- *Initial and final substrings of the words.* We compute the initial and final substrings of different length, from three to five characters.
- *Number and percentage of consonants, vowels and other characters.* We compute the number and the percentage of consonants, vowels and other characters (i.e., hyphen, apostrophe, comas).
- *Classical readability scores.* We apply two classical readability measures: Flesch (Flesch, 1948) and its variant Flesch-Kincaid (Kincaid et al., 1975). Such measures are typically used for evaluating the difficulty level of a text. They exploit surface characteristics of words (number of characters and/or syllables) and normalize these values with specifically designed coefficients.

5.1.2 FastText word embeddings usage

Currently, *word embedding vectors* (Mikolov et al., 2013) (or *word vector representations*) are used in the most of state-of-the-art methods for various NLP tasks (*Repository to track the progress in Natural Language Processing (NLP)*). Usually, word embeddings are pre-trained on the giant corpora of natural texts such as Google News, Wikipedia texts in an unsupervised manner to predict the context of the target words. They exploit the distributional hypothesis that semantically close words are next to each other in the sentence and that semantically close words share similar contexts.

FastText word embeddings (Bojanowski et al., 2016) is a good candidate as features for words difficulty detection task because they are able to use words' morphological information and generalize over it. The fact that word embeddings capture context and morphological information leads to the hypothesis that incorporating this information as features will improve classification accuracy for our specific problem. FastText embedding vectors are the sum of character n-gram representations, so that they could be generated even for unknown words. Nevertheless, being trained on Wikipedia texts the portion of known words from our dataset for current FastText embeddings is quite big. According to our analysis, 44.26% (13,118 out of 29,641) medical words in the dataset and 56.00% (16,598 out of 29,641) lowercased

medical words in the dataset were used for training of the currently published FastText[3] model for French.

5.1.3 Word embeddings from RNN

[3] https://fasttext.cc

Chapter 6

Experiments

We conducted a number of experiments to detect the words' understandability and generalization properties of resulting models. The quality of the applied classification algorithms was evaluated using four standard measures: accuracy A, precision P, recall R and F1-measure F. These scores are weighted average for 1-vs-rest binary classifiers for each of three classes, described in 4.2. Such evaluation of models allows to measure the ability of a chosen methodology (a feature set and a classification method) to distinguish understandable and non-understandable words in an unbalanced dataset.

6.1 Reproduction of previous results

In (Grabar, Hamon, and Amiot, 2014) the classification methods were obtained using WEKA [1] - a collection of machine learning algorithms for data mining tasks implemented on Java. In our research as a tool to conduct experiments we used Python, because it is easy to use and there are a lot of stable third-party Python libraries that make Python convenient for research. In order to ensure the consistency of experiments in this work and in (Grabar, Hamon, and Amiot, 2014), firstly, we reproduced the WEKA results using pre-computed standard set of features from (Grabar, Hamon, and Amiot, 2014) and J48 classification algorithm based on Decision Tree (DT) - a WEKA implementation of C4.5 described in section (Quinlan, 1993). Our results perfectly match with ones presented in paper. Secondly, we developed a solution based on DT classifier from well-known scikit-learn library[2]. At this step we got 0.85-1.41 lower F scores for scikit-learn compared to WEKA results (Table 6.1).

annotator \method	*Results from paper (Grabar, Hamon, and Amiot, 2014)*	*WEKA J48*	*Python Decision trees (10-fold CV, with shuffle)*
$A1$	80.6	80.5	79.8
$A2$	81.4	80.9	80.0
$A3$	84.5	84.5	83.2

TABLE 6.1: Comparison of various implementations for decision tree classifier on three datasets (A1, A2, A3) in user-in vocabulary-out cross-validation

[1] https://www.cs.waikato.ac.nz/ml/weka/
[2] http://scikit-learn.org

Since the input features were identical for WEKA and scikit-learn frameworks, we decided that this small degradation of quality is caused by the difference in implementations of decision tree classifiers in these frameworks. And so, in all subsequent experiments we will use the scikit-learn results reproduction because of it's convenience for comparison of experiments' results.

6.2 Selection of classifier

Here will be shortly introduced results on using different classifiers and that DT was the best.

6.3 Experiments with cross-validation settings

6.3.1 User-in vocabulary-out cross-validation

These experiments also follows the scenario from (Grabar, Hamon, and Amiot, 2014). The cross-validation is done on each dataset (i.e. each user's annotation) separately. The goal of these experiments is to measure the ability of the method to generalize class recognition on the *known user*and his known manner to annotate words (that is, his understanding of the meaning of medical words) for *unknown words*.

We carried out the experiments using (i) the standard features only, (ii) the FastText word embeddings only and (iii) their combination. Experiments with isolated FastText word embeddings as features and the data from three annotators resulted in poor F-scores (Table 6.2), that can be treated that contextual information which is dominant in the word embeddings is not enough to define the word understandability. Adding the FastText word embeddings to the standard feature set resulted in up to 1% higher F-score due to higher Precision (up to 1.8%), meaning that contextual information slightly impacts on the understandability of a word by a given person.

Train user	*Test user*	*Standard features only*				*Embeddings only*				*Standard features + FastText word embeddings*			
		A	*P*	*R*	*F*	*A*	*P*	*R*	*F*	*A*	*P*	*R*	*F*
*A*1	*A*1	**82.5**	77.2	**82.5**	79.8	72.5	67	72.5	69.3	82.4	**79**	82.4	**80.2**
*A*2	*A*2	**82**	78.9	**82**	80	73.5	69.9	73.5	71.3	81.9	**79.5**	81.9	**80.3**
*A*3	*A*3	85.5	81.2	85.5	83.2	74.9	70.4	74.9	72.3	**85.9**	**83**	**85.9**	**84.2**

TABLE 6.2: Experiments on user-in vocabulary-out cross-validation

6.3.2 User-out vocabulary-in cross-validation

In this experiment, we learn from all the annotations of one user and then test the model on annotations of another user. In this setting, we measure the ability of the classifier to generalize on all known words, but for unknown users (Table 6.3). This scenario is realistic to a real-world situation: the reference annotations can be obtained only from a couple of users, presumably representing the overall population, but not from all the possible users. Yet, it is necessary to predict the familiarity of medical words for all the potential users even if they did not participate in the annotations.

Train user	*Test user*	*Standard features only*				*Embeddings only*				*Standard features + FastText word embeddings*			
		A	*P*	*R*	*F*	*A*	*P*	*R*	*F*	*A*	*P*	*R*	*F*
A1	*A2*	81.7	78.6	81.7	80.1	74	70.3	74	71.2	**84.2**	**82**	**84.2**	**82.8**
A1	*A3*	85	81.2	85	83	75.4	70.7	75.4	72.6	**87.6**	**84.9**	**87.6**	**85.9**
A2	*A1*	82.2	77	82.2	79.1	72.8	67.3	72.8	69.6	**83.9**	**80.2**	**83.9**	**81.1**
A2	*A3*	85.4	81.1	85.4	83	75.3	71.1	75.3	73	**86.8**	**83.5**	**86.8**	**84.7**
A3	*A1*	82.8	77.4	82.8	79.7	72.7	67.1	72.7	69.4	**84.9**	**81.3**	**84.9**	**82.4**
A3	*A2*	82.2	79	82.2	80.2	74.1	70.4	74.1	71.6	**84.2**	**82.1**	**84.2**	**82.8**

TABLE 6.3: Experiments on user-out vocabulary-in cross-validation

In these experiments we got a significant improvement of combined features in comparison to the standard features. When knowledge of words understandability of one user is used to predict it for another user, adding the FastText word embeddings provides up to 2.9 better F-score. Notice that used separately, standard features and embeddings shows similar performance as in user-in vocabulary-out cross-validation (Table 6.3). Our hypothesis is that there exists a robust nonlinear dependency between some subsets of standard features and subword-level components of FastText word embeddings. Testing this hypothesis is the topic of our further research.

6.3.3 User-out vocabulary-out cross-validation

In this experiment, we take (k-1) folds of data from one user for training and use k-th fold for testing from the remaining user. In this case, we measure the ability of the method to generalize both on *unknown users* and *unknown vocabulary*.

The cross-validation setting is now the most strict and knowledge of words understandability of one user is used to predict whether another user will understand other medical words. In these experiments, embeddings provide approximately 0.5% higher F-score in case of learning on users A1 and A3 (Table 6.4). When learning on user A2, embeddings decrease F by 0.5, which means that annotations and health literacy of user A2 are different from users A1 and A3. It seems that adding embeddings makes overfitting of machine learning model to the dataset. As a result, tests on other "kind of word understandability" and on combined features are less successful compared to using standard features only for learning. This may be due to the lack of systematicity in annotations of A2.

Train user	*Test user*	*Standard features only*				*Embeddings only*				*Standard features + FastText word embeddings*			
		A	*P*	*R*	*F*	*A*	*P*	*R*	*F*	*A*	*P*	*R*	*F*
A1	*A2*	81.7	78.6	81.7	80.1	73.6	69.9	73.6	71.3	**81.8**	**79.8**	**81.8**	**80.6**
A1	*A3*	**85**	81.2	**85**	83	74.8	70.4	74.8	72.4	84.9	**82.2**	84.9	**83.4**
A2	*A1*	**82.2**	76.9	**82.2**	**79.1**	72.5	66.9	72.5	69.3	81.7	**77.5**	81.7	**79.1**
A2	*A3*	**85.3**	81	**85.3**	**83**	75.1	70.7	75.1	72.7	84.4	**81.3**	84.4	82.5
A3	*A2*	**82.7**	77.3	**82.7**	79.7	72.5	66.9	72.5	69.2	82.6	**78.9**	82.6	**80.2**
A3	*A3*	82.1	79	82.1	80.1	73.8	70.2	73.8	71.4	**82.2**	**80**	**82.2**	**80.7**

TABLE 6.4: Experiments on user-out vocabulary-out cross-validation

6.4 Generalizability study

In the previous experiments we concentrated on three annotators' data to be consistent with the research in paper (Grabar, Hamon, and Amiot, 2014). To study better generalizability of models for words' understandability detection we included 4 more annotators in an experiment.

In this part we concentrated on the user-out vocabulary-in cross-validation scenario as the most realistic one. Here understanding of generalizability quality is crucial for usage of the model in real world client-doctor relationship.

The results obtained are presented in Table 6.5. The first two columns indicate the annotators. Data provided by each annotator are used for training the classifier (first column). The model generated is then tested on data from all the annotators including the reference annotator (second column). Three sets of such experiments are performed, depending on features exploited: standard features, word embeddings, and combination of all the features available. Each experiment is evaluated with several measures: *P* Precision, *R* Recall, *F* F-measure to evaluate the efficiency in prediction which medical words are understandable or not understandable for a given annotator.

We can do several observations on these results. Features used shows an impact on the results obtained. Thus, standard features usually show better results than embeddings. One explanation is that standard features include 24 individual features covering different aspects of linguistic and non-linguistic description of words, while word embeddings rely only on distribution of words and their similarity. Yet, combination of all the features (standard and embeddings) usually improves overall results, sometimes going to up to 2.9 improvement of F-measure. Our hypothesis is that there exists a robust nonlinear dependency between some subsets of standard features and subword-level components of FastText word embeddings. Testing this hypothesis is the topic of our further research.

Recall values are always higher than Precision values. In each set of experiments, the best results are not obtained when the model of a given annotator is applied to own data. For instance, the *O1* model provides better results when tested on data from annotators *O2*, *O3* and *A8*. Similarly, the *A7* model shows better results when applied to data from annotators *O1*, *O2*, *O3* and *A8*. This is an important issue because it shows that the models acquired from one annotator can be successfully generalized over other annotators.

Besides, it seems that the annotators form two clusters according to the classification of difficult medical words: one cluster with four annotators (*O1*, *O2*, *O3*, *A8*) and one cluster with three annotators (*A1*, *A2*, *A7*). This issue may be related to the health literacy of annotators. This may indicate that the annotation models can be shared by people with similar skills and knowledge. Yet, to confirm this hypothesis, it is necessary to define the level of health literacy of annotators. This task is rather difficult because there is no existing tests created for computing the health literacy level for French-speaking healthy people. Another hypothesis is that some models may be better generalizable than other models. This hypothesis must also be verified with additional experiments;

Another important point is that, while the annotations go forward, the annotators usually show *learning* progress in decoding the morphological structure of terms and their understanding (Grabar and Hamon, 2017). This progress is not taken into account in the current experiments.

Train annotator	*Test annotator*	*Standard features*			*Embeddings*			*Standard features embeddings*		
		P	R	F	P	R	F	P	R	F
O1	O1	77.2	82.5	79.7	67.0	72.5	69.3	79.0	82.4	80.2
O1	O2	78.6	81.7	80.1	70.3	74.0	71.2	82.0	84.2	82.8
O1	O3	81.2	85.0	83.0	70.7	75.4	72.6	84.9	87.6	85.9
O1	A1	71.0	74.7	71.2	62.1	63.8	58.8	74.1	75.4	72.2
O1	A2	70.6	78.4	74.0	61.9	68.5	63.3	75.0	80.1	76.2
O1	A7	72.6	77.5	74.2	63.0	66.6	61.9	76.2	78.9	75.8
O1	A8	82.3	84.9	83.5	73.1	76.8	74.5	85.7	87.8	86.6
O2	O1	77.0	82.2	79.1	67.3	72.8	69.6	80.2	83.9	81.1
O2	O2	78.9	82.0	80.0	69.9	73.5	71.3	79.5	81.9	80.3
O2	O3	81.1	85.4	83.0	71.1	75.3	73.0	83.5	86.8	84.7
O2	A1	71.1	72.1	68.2	61.7	64.5	60.2	74.0	75.1	71.5
O2	A2	70.8	77.3	72.7	61.8	68.9	64.2	76.0	79.8	75.5
O2	A7	72.7	75.6	71.8	62.6	67.0	62.8	75.9	78.3	74.9
O2	A8	83.0	86.2	84.4	73.7	77.1	75.3	85.4	88.2	86.7
O3	O1	77.4	82.8	79.7	67.1	72.7	69.4	81.3	84.9	82.4
O3	O2	79.0	82.2	80.2	70.4	74.1	71.6	82.1	84.2	82.8
O3	O3	81.2	85.5	83.2	70.4	74.9	72.3	83.0	85.9	84.2
O3	A1	71.8	73.3	69.5	61.7	64.1	59.6	75.1	75.4	72.1
O3	A2	71.2	78.0	73.5	61.8	68.7	63.9	76.8	80.2	76.3
O3	A7	73.2	76.5	72.9	62.4	66.6	62.2	77.2	78.8	75.8
O3	A8	82.6	85.8	84.1	73.7	77.2	75.2	86.0	88.0	86.9
A1	O1	77.2	82.5	79.8	66.5	67.9	66.6	76.9	79.5	77.6
A1	O2	78.6	81.6	80.1	69.2	69.0	68.5	78.8	79.6	78.9
A1	O3	81.2	84.9	82.9	70.7	69.6	69.2	81.8	82.0	81.0
A1	A1	70.9	74.7	71.3	59.4	64.6	61.8	72.4	75.1	72.9
A1	A2	70.5	78.3	74.0	60.6	66.4	63.2	73.7	78.6	75.0
A1	A7	72.6	77.5	74.2	61.3	66.1	63.6	75.1	79.2	76.5
A1	A8	82.2	84.8	83.5	72.3	70.4	70.4	81.5	81.0	80.5
A2	O1	77.3	82.6	79.8	67.2	72.6	69.6	81.0	82.8	81.8
A2	O2	78.6	81.6	80.1	70.4	74.0	71.9	82.0	82.0	82.0
A2	O3	81.2	84.9	83.0	71.0	75.2	73.0	84.9	85.4	85.1
A2	A1	70.9	74.6	71.2	61.5	64.6	60.4	76.5	76.5	74.7
A2	A2	70.6	78.4	74.0	61.2	68.4	63.7	74.7	77.8	75.6
A2	A7	72.6	77.5	74.2	62.4	67.0	63.0	77.6	78.9	77.3
A2	A8	82.2	84.8	83.4	73.8	77.0	75.3	85.6	85.3	85.4
A7	O1	77.1	82.5	79.7	67.6	73.2	69.9	79.4	81.9	80.3
A7	O2	78.5	81.6	80.0	70.6	74.2	71.8	80.6	81.4	80.9
A7	O3	81.0	84.9	82.9	71.3	75.7	73.3	83.1	83.8	83.0
A7	A1	71.0	74.4	70.9	62.1	64.8	60.3	75.8	78.0	75.7
A7	A2	70.5	78.2	73.8	62.0	69.1	64.3	75.3	79.6	76.5
A7	A7	72.6	77.4	74.0	62.2	67.0	63.1	74.5	77.5	75.3
A7	A8	81.9	84.7	83.3	73.7	77.2	75.3	82.8	82.7	82.4
A8	O1	77.0	82.4	79.6	67.2	72.7	69.6	80.8	84.4	81.7
A8	O2	78.4	81.5	79.8	70.4	74.0	71.7	82.0	84.7	83.0
A8	O3	80.9	84.9	82.8	71.0	75.2	72.9	84.7	87.6	85.6
A8	A1	71.0	74.2	70.7	61.4	64.3	60.0	73.7	75.0	71.5
A8	A2	70.4	78.1	73.7	61.7	68.8	64.1	75.0	80.1	75.9
A8	A7	72.6	77.2	73.7	62.2	66.6	62.5	75.7	78.2	74.9
A8	A8	81.9	84.9	83.4	73.6	77.0	75.1	84.2	86.5	85.2

TABLE 6.5: Experiments on portability of models from one user to another

Chapter 7

Conclusions

7.1 Contribution

We proposed to address the detection of medical words which understanding may be difficult for non-specialized users of the medical area. We exploit for this machine learning algorithms, reference data from seven annotators, and several sets of NLP features: standard features (syntactic information, reference lexica, frequency, etc.), distributional features (word embeddings), and their combination.

Our results provide several indications. Hence, the combination of all features is the most efficient. Concerning the generalization, we propose to learn model on a given annotator and then to apply it to data obtained from other annotators. This set of experiments indicate that models provide better results when tested on data from other annotators. We consider this to be a positive issue because it is important to be able to generalize annotations provided by a set of users on the whole population. Yet, these results may point out that the users should be apprehended through their health literacy, while currently there is no available tests for measuring it in French-language healthy people.

7.2 Future work

We have several directions for future work. We currently use existing pre-trained word embeddings. Yet, we assume that their training on medical data may improve their impact on the categorization results. We also plan to implement and test other deep learning/neural networks/NLP methods which use the morphological information of words, such as character-level recurrent neural networks and character embeddings together with 1D convolutions. Indeed, when language data present stable patterns, which is the case in the medical field, processing of sub-word strings may help for the generalization over new and unseen words. As we presented above, this is one of the current limitations of our work.

Bibliography

Amoia, M and M Romanelli (2012). "SB: mmSystem - Using Decompositional Semantics for Lexical Simplification". In: **SEM 2012*. Montréal, Canada, pp. 482–486. URL: http://www.aclweb.org/anthology/S12-1067.

Bojanowski, P. et al. (2016). *Enriching word vectors with subword information*. arXiv preprint arXiv:1607.04606.

Borst, A et al. (2008). "Lexically based distinction of readability levels of health documents". In: *MIE 2008*. Poster.

Brigo, F et al. (2015). "Clearly written, easily comprehended ? The readability of websites providing information on epilepsy". In: *Epilepsy & Behavior* 44, pp. 35–39.

Chmielik, J and N Grabar (2011). "Détection de la spécialisation scientifique et technique des documents biomédicaux grâce aux informations morphologiques". In: *TAL* 51.2, pp. 151–179.

Clercq, Orphée De et al. (2014). "Using the crowd for readability prediction". In: *Natural Language Engineering* 20, pp. 293–325.

Côté, Roger A. et al. (1993). *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*. Northfield: College of American Pathologists.

Eysenbach, Gunther (2007). "Poverty, Human Development, and the Role of eHealth". In: *J Med Internet Res* 9.4, pp. 34–4.

Flesch, R (1948). "A new readability yardstick". In: *Journ Appl Psychol* 23, pp. 221–233.

François, T and C Fairon (2013). "Les apports du TAL à la lisibilité du français langue étrangère". In: *TAL* 54.1, pp. 171–202.

Gala, N, T François, and C Fairon (2013). "Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons". In: *eLEX-2013*.

Goeuriot, L, N Grabar, and B Daille (2008). "Characterization of scientific and popular science discourse in French, Japanese and Russian". In: *LREC*.

Grabar, N, S Krivine, and MC Jaulent (2007). "Classification of Health Webpages as Expert and Non Expert with a Reduced Set of Cross-language Features". In: *Ann Symp Am Med Inform Assoc (AMIA)*, pp. 284–288.

Grabar, Natalia, Emmanuel Farce, and Laurent Sparrow (2018). "Study of readability of health documents with eye-tracking and machine learning approaches". In: *Int Conf on Healthcare Informatics (ICHI)*. Poster, pp. 1–2.

Grabar, Natalia and Thierry Hamon (2016). "A large rated lexicon with French medical words". In: *LREC (Language Resources and Evaluation Conference)*, pp. 1–12.

— (2017). "Understanding of unknown medical words". In: *Proceedings of the Biomedical NLP Workshop associated with RANLP 2017*. Varna, Bulgaria: INCOMA Ltd., pp. 32–41. DOI: 10.26615/978-954-452-044-1_005. URL: https://doi.org/10.26615/978-954-452-044-1_005.

Grabar, Natalia, Thierry Hamon, and Dany Amiot (2014). "Automatic diagnosis of understanding of medical words". In: *EACL PITR Workshop*, pp. 11–20.

Gunning, R (1973). *The art of clear writing*. New York, NY: McGraw Hill.

Jauhar, SK and L Specia (2012). "UOW-SHEF: SimpLex – Lexical Simplicity Ranking based on Contextual and Psycholinguistic Features". In: **SEM 2012*. Montréal, Canada, pp. 477–481. URL: http://www.aclweb.org/anthology/S12-1066.

Johannsen, A et al. (2012). "EMNLP@CPH: Is frequency all there is to simplicity?" In: **SEM 2012*. Montréal, Canada, pp. 408–412. URL: http://www.aclweb.org/anthology/S12-1054.

Jucks, R and R Bromme (2007). "Choice of words in doctor-patient communication: an analysis of health-related internet sites". In: *Health Commun* 21.3, pp. 267–77.

Kincaid, JP et al. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Tech. rep. Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

Kokkinakis, D and M Toporowska Gronostaj (2006). "Comparing Lay and Professional Language in Cardiovascular Disorders Corpora". In: *WSEAS Transactions on BIOLOGY and BIOMEDICINE*. Ed. by Australia Pham T. James Cook University, pp. 429–437.

Ligozat, AL et al. (2012). "ANNLOR: A Naïve Notation-system for Lexical Outputs Ranking". In: **SEM 2012*, pp. 487–492.

Mcgray, A (2005). "Promoting Health Literacy". In: *J of Am Med Infor Ass* 12, pp. 152–163.

Mikolov, T et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *NIPS*.

Miller, T et al. (2007). "A Classifier to Evaluate Language Specificity of Medical Documents". In: *HICSS*, pp. 134–140.

Namer, F (2000). "FLEMM : un analyseur flexionnel du français à base de règles". In: *Traitement automatique des langues (TAL)* 41.2, pp. 523–547.

Namer, Fiammetta and Pierre Zweigenbaum (2004). "Acquiring meaning for French medical terminology: contribution of morphosemantics". In: *Ann Symp Am Med Inform Assoc (AMIA)*. San-Francisco.

Oregon Practice Center (2008). *Barriers and Drivers of Health Information Technology Use for the Elderly, Chronically Ill, and Underserved*. Tech. rep. Agency for healthcare research and quality. Oregon Evidence-based Practice Center.

Patel, V, T Branch, and J Arocha (2002). "Errors in interpreting quantities as procedures : The case of pharmaceutical labels". In: *Int Journ Med Inform* 65.3, pp. 193–211.

Poprat, M, K Markó, and U Hahn (2006). "A Language Classifier that Automatically Divides Medical Documents for Experts and Health Care Consumers". In: *Int Congress of the European Federation for Medical Informatics*. Maastricht, pp. 503–508.

Quinlan, JR (1993). *C4.5 Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Repository to track the progress in Natural Language Processing (NLP). https://nlpprogress.com. Last accessed 30 October 2018.

Schmid, H (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees". In: *Int Conf on New Methods in Language Processing*, pp. 44–49.

Sinha, R (2012). "UNT-SimpRank: Systems for Lexical Simplification Ranking". In: **SEM 2012*, pp. 493–496.

Specia, L, SK Jauhar, and R Mihalcea (2012). "SemEval-2012 Task 1: English Lexical Simplification". In: **SEM 2012*, pp. 347–355.

Tran, TM et al. (2009). "Internet et soins : un tiers invisible dans la relation médecin/patient ?" In: *Ethica Clinica* 53, pp. 34–43.

Vander Stichele, RH (1999). "Promises for a measurement breakthrough". In: *Drug regimen compliance. Issues in clinical trials and patient management*. Ed. by John Wiley & Sons. JM Metry and UA Meyer, pp. 71–83.

Williams, MV et al. (1995). "Inadequate functional health literacy among patients at two public hospitals". In: *JAMA* 274.21, pp. 1677–1682.

Yaneva, V, I Temnikova, and R Mitkov (2015). "Accessible texts for autism: An eye-tracking study". In: *Int ACM SIGACCESS Conference on Computers & Accessibility*. Ed. by ACM, pp. 49–57.

Zeng-Treiler, Q et al. (2007). "Text characteristics of clinical reports and their implications for the readability of personal health records". In: *MEDINFO*. Brisbane, Australia, pp. 1117–1121.

Zheng, W, E Milios, and C Watters (2002). "Filtering for medical news items using a machine learning approach". In: *Ann Symp Am Med Inform Assoc (AMIA)*, pp. 949–53.