RNN Embeddings for Identifying Difficult to Understand Medical Words

Hanna Pylieva[1], Artem Chernodub[1, 2], Natalia Grabar[3], and Thierry Hamon[4,5]

[1]Faculty of Applied Sciences, Ukrainian Catholic University, Lviv, Ukraine
{pylieva,chernodub}@ucu.edu.ua
[2]Grammarly, Kyiv, Ukraine
[3]CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000, Lille, France
natalia.grabar@univ-lille.fr
[4]LIMSI, CNRS, Université Paris-Saclay, F-91405, Orsay, France
hamon@limsi.fr
[5]Université Paris 13, Sorbonne Paris Cité, F-93430, Villetaneuse, France

Abstract

Patients and their families often require a better understanding of medical information provided by doctors. We currently address this issue by improving the identification of difficult to understand medical words. We introduce novel embeddings received from RNN - FrnnMUTE (French RNN Medical Understandability Text Embeddings) which allow to reach up to 87.0 F1 score in identification of difficult words. We also note that adding pre-trained FastText word embeddings to the feature set substantially improves the performance of the model which classifies words according to their difficulty. We study the generalizability of different models through three cross-validation scenarios which allow testing classifiers in real-world conditions: understanding of medical words by new users, and classification of new unseen words by the automatic models. The RNN - FrnnMUTE embeddings and the categorization code are being made available for the research.

1 Introduction

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, and human anatomy. 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 by experts in the medical area texts. As in any specialized areas, two main kinds of users exist in the medical area: experts of the domain, i.e. medical doctors, both researchers or practitioners; consumers of the healthcare process, i.e. patients and their relatives. The latter usually do not have expert knowledge in the medical domain, while it is important that they understand the purpose and issues of their healthcare process.

The existing literature provides several studies dedicated to the understanding of medical notions and terms by non-expert users, and of how the level of health literacy of patients impacts on a successful healthcare process (McCray, 2005; Eysenbach, 2007), as indeed it is quite common that patients and their relatives must face technical health documents and information. These observations provide the main motivation for our work. Hence, we address the need of non-specialized users in the medical domain to understand medical and health information.

In this paper, we propose to apply deep learning techniques to improve identification of readability and understandability of medical words by non-expert users. In particular, we will tackle the word categorization task and compare the performance of classification model on different feature sets: standard linguistic and non-linguistic features described in section 4, features obtained using different deep learning approaches, and the combinations of all of them. We also investigate how different feature sets perform with three different cross-validation settings, described in section 5. The medical data used in this work are in French. Three human annotators participated in the creation of the reference data (specifying the understandability of words).

2 Related Work

Related work is globally positioned in the text simplification task which involves the detection of complex contents in documents and their adaptation for the target population. We are also interested in the first aspect with additional constraints:

detection and diagnosis of technical contents in texts from medical domain.

Traditional readability measures rely on two main factors: the familiarity of semantic units such as words or phrases, and the complexity of syntax. To make these measures straightforward for applications, some simplifying assumptions were used. As a result, final formulas mostly rely on the number of letters and/or of syllables a word contains and on linear regression models (Flesch, 1948; Gunning, 1973). While such readability measures are easy to compute, they are based on shallow characteristics of text, ignoring deeper levels of text processing which are important factors in readability, such as cohesion, syntactic ambiguity, rhetorical organization, and propositional density (Collins-Thompson, 2014). Moreover, traditional readability measures were demonstrated to be unreliable for non-traditional documents (Si and P. Callan, 2001). As a result of such limitations and due to the recent growth of computational and data resources, the focus of NLP researchers moved to *computational* readability measurements, which rely on the use of machine learning algorithms on richer linguistic features (Malmasi et al., 2016; Ronzano et al., 2016; Bingel et al., 2016).

Not so much effort has been devoted to the exploitation of NLP potential in the measurement of readability of medical texts. In the biomedical domain, as well as in general language, the readability assessment is currently approached as a classification task. The difference is that in the former a much smaller variety of features has been tested: a combination of classical readability formulas with medical terminologies (Kokkinakis and Toporowska Gronostaj, 2006); n-grams of characters (Poprat et al., 2006); stylistic (Grabar et al., 2007) or discursive (Goeuriot et al., 2008) features; morphological features (Chmielik and Grabar, 2011); combinations of different features from those listed above (Zeng-Treiler et al., 2007). Among the recent experiments dedicated to readability study in the medical domain are, for example, manual rating of medical words (Zheng et al., 2002), automatic rating of medical words on the basis of their presence in different vocabularies (Borst et al., 2008), exploitation of machine learning approach with various features (Grabar et al., 2014). The last experiment achieved up to 85.0 F-score on individual annotations.

Due to the recent significant advance in readability study in general language and relatively slow progress with the task in the medical area, there is a great potential of experimenting with the machine learning-based approaches on medical texts. This fact motivated us for choosing this kind of methodology.

3 Materials

For the experiments, we used the publicly available set of words with annotations[1]. The process of words collection and annotation is briefly described below.

3.1 Linguistic data description

The set of required biomedical terms was obtained from the French part of Snomed International[2] (Côté et al., 1993). Snomed Int 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 word understandability study, five axes related to the main medical notions were chosen: disorders, abnormalities, procedures, functions, and anatomy. These categories are assumed to be faced frequently by layman. In contrast, chemical products and living organisms are excluded because they mainly correspond to Latin borrowings and are typically non-understandable by laypeople.

The 104,649 selected terms were then processed. First, they were tokenized, POS-tagged and lemmatized using TreeTagger (Schmid, 1994). Then the lemmatization was checked with FLEMM (Namer, 2000). After that we received 29,641 unique words. For instance, the term *trisulfure d'hydrogène* provided three words (*trisulfure, de, hydrogène*). The final dataset contains compound words which contain several bases (*abdominoplastie* (abdominoplasty), *dermabrasion* (dermabrasion)), constructed words which contain one base and at least one affix (*lipoïde* (lipoid), *cardiaque* (cardiac)), simple words which contain one base, no affixes and possibly infections when the lemmatization fails (*acné* (acne), *fragment* (fragment)).

[1] http://natalia.grabar.free.fr/resources.php#rated

[2] https://esante.gouv.fr/terminologie-snomed-35vf

Annotators / Categories	*Cat1*	*Cat2*	*Cat3*	*Total*
O1 (%)	8,099 (28)	1,895 (6)	19,647 (66)	29,641
O2 (%)	8,625 (29)	1,062 (4)	19,954 (67)	29,641
O3 (%)	7,529 (25)	1,431 (5)	20,681 (70)	29,641

Table 1: Number (and percentage) of words assigned to reference categories by seven annotators (O1, O2, O3).

3.2 Annotation process

The set of 29,641 unique words was annotated by three French speakers, 25-40-year-old, without medical training, without specific medical problems, but with the linguistic background. The annotators were expected to represent the average knowledge of medical words among the population as a whole. They were presented with a list of terms and asked to assign each word to one of the three categories: (Cat1) *I can understand the word*; (Cat2) *I am not sure about the meaning of the word*; (Cat3) *I cannot understand the word*. The annotators were asked not to use dictionaries during the annotation process. The interannotator agreement shows substantial agreement: Fleiss' Kappa 0.735 and Cohen's Kappa 0.736. This is a very good result, especially when working with linguistic data for which the agreement is usually difficult to obtain. The annotation results are represented in Table 1.

4 Method

We aim to categorize medical words according to whether they can be understood or not by non-specialized people, using features obtained with NLP tools and with deep learning methods. The manual annotations of these words described in the previous section provide the reference data. The proposed method includes calculation of NLP features associated with the annotated words, training machine learning models for word classification, and evaluation of classification using cross-validation.

4.1 Feature sets

We distinguish and use two kinds of features: standard features provided by the NLP analysis of words, and features issued from existing or specifically trained word embeddings. These two types of features are first opposed and then combined.

4.1.1 Standard NLP features

The standard NLP features 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 enriched by FLEMM (Namer, 2000).
- *Presence of words in reference lexica.* Two reference lexica of the French language were exploited: TLFi[3] and *lexique.org*[4]. 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, and almost 35,000 lemmas.
- *Frequency of words through a non specialized search engine.* Each word were queried on Google to find out the frequency of the word on the web.
- *Frequency of words in the medical terminology.* The frequency of words in the medical terminology Snomed Int corresponds to the number of different terms containing a given word.
- *Number and types of semantic categories associated to words.* The information on the semantic categories of Snomed Int was exploited.
- *Length of words in number of their characters and syllables.* For each word, the number of its characters and syllables was computed.
- *Number of bases and affixes.* Each lemma was 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

[3] http://www.atilf.fr/
[4] http://www.lexique.org/

and it provides also semantic explanation of the analyzed lexemes. The morphological decomposition information (number of affixes and bases) was exploited. For instance, *hématomètre* (*haemometer*) is analyzed and decomposed into two basis (*hémato* meaning *blood* and *mètre* meaning *measure*, while *myélite* (*myelitis*) is decomposed into *myél* meaning *marrow* and *ite* meaning *inflammation*.

- *Initial and final substrings of the words.* Initial and final substrings of different length, from three to five characters, were computed.
- *Number and percentage of consonants, vowels and other characters.* The number and the percentage of consonants, vowels and other characters (i.e., hyphen, apostrophe, comas) was computed.
- *Classical readability scores.* Two classical readability measures were applied: 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.

4.1.2 FastText word embeddings usage.

FastText word embeddings (Bojanowski et al., 2017) is a good candidate feature for the detection of word difficulty because they are able to use the morphological information of words and generalize over it. Since the word embeddings capture context and morphological information, we assume that using them as features will improve classification accuracy for our specific problem.

We note that FastText word embeddings trained on Wikipedia and Common Crawl[5] texts have an important part of words from our dataset. According to our analysis, the currently published FastText[6] model for French contains 44.26% (13,118 out of 29,641) medical words from our dataset and up to 56.00% (16,598 out of 29,641) lowercased medical words from our dataset.

4.1.3 French RNN Medical Understandability Text Embeddings (FrnnMUTE).

According to the general functionality of RNNs, the final hidden state aggregates the information about all input sequence. This idea is frequently used to receive hidden representations of sequences. Sequence-to-sequence model is a well-known example of how this idea works in practice (Sutskever et al., 2014). Such models consist of two parts: an *encoder* is an RNN which encodes the input sequence into a representation in hidden space (which is also called *thought vector*), and a *decoder* which generates a new sequence out of the hidden representations.

We used this idea for representing words from our dataset. To receive words representations from an RNN, we first trained it to classify words based on labels by one annotator (we chose $O1$), then for each word we collect values of the last hidden state of the RNN and use this vector as features during the detection of words understandability for different users (or annotators). Train/test split was 70%/30% of randomly shuffled samples.

As a direct classifier, we trained a character-level RNN using PyTorch framework[7] and one GPU Tesla K80. We lowercased all words, lemmatized them and substituted all Unicode symbols with their ASCII analogs. We tested several RNN architectures and hyperparameter sets. The best performance was reached with a model consisting of two unidirectional long short-term memory (LSTM) units, each with 50 hidden units. The dropout of the model is 0.7. The input size is 57 as the number of unique characters in lowercased and converted to ASCII input words. The output size is 3 as the number of classes in our data. This model reached the best performance on the eighth epoch with $F1 = 78.94$ and $accuracy = 81.21\%$ on development set. Using this model we received 50-dimensional word representations which we called FrnnMUTE (French RNN Medical Understandability Text Embeddings).

5 Experiments and Results

We study the impact of adding words embeddings as features for identifying difficult for understanding words. First, we observe how FastText word embeddings influence the quality of classification in different cross-validation scenarios. Then, we study how FrnnMUTE used as features impact on classification quality in all the same cross-validation scenarios. The quality of the classifications is evaluated using four standard *macroaveraging* (Sebastiani, 2002) measures: accuracy A,

[5] http://commoncrawl.org/
[6] https://fasttext.cc
[7] https://pytorch.org/

precision P, recall R and F1-measure F.

5.1 Cross-validation scenarios

For a thorough study of generalization abilities of the classification models, we propose to consider three distinct cross-validation scenarios based on different combinations of users and vocabulary in train and test sets.

5.1.1 User-in vocabulary-out cross-validation

The cross-validation is performed on each dataset (i.e., each user annotation) separately. We aim to measure the ability of the classification model to generalize class recognition on the *known user* and to predict annotations for *unknown words*. From the practical perspective, *user-in* means learning the profile of a user. Hence, a model trained by such scenario represents the word understanding or knowledge of the annotator.

The experiments use (i) the standard features only, (ii) the FastText word embeddings only and (iii) their combination. The experiments with isolated FastText word embeddings as features resulted in poor F1 scores (Table 2), that can be explained by the fact that contextual information, which is dominant in these word embeddings, is not enough to define the word understandability. Adding the FastText word embeddings to the standard feature set resulted in up to 1.0 higher F1 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.

5.1.2 User-out vocabulary-in cross-validation

We then learn from all the annotations of one user and then test the model on annotations of another user. Thereby, in such a setting, we measure the ability of the classifier to generalize on all known words, but for unknown users. 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. In this scenario, the model learns the profile of a user and we want to identify whether a new user has the same profile as another user. Then it can be used for identification of not understandable words for the new users.

These experiments show a substantial improvement of combined features in comparison to the standard features (Table 3). 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 F1 score. Used separately, standard features and embeddings show similar performance as in user-in vocabulary-out cross-validation (Table 2). We assume 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 future work.

5.1.3 User-out vocabulary-out cross-validation

Finally, we consider (k-1) folds of data from one user for training and use k-th fold for testing from the remaining user. We aim to measure the ability of the method to generalize both on *unknown users* and *unknown vocabulary*. This experiment should be helpful in identifying the number of words needed for determining whether the profile of one user is the same as profile of other users in case the model achieves good performance.

In these experiments, FastText word embeddings provide approximately 0.5% higher F1 score in case of learning on users O1 and O3 (Table 4). When learning on user O2, embeddings decrease F by 0.5, which means that annotations and health literacy of user O2 are different from users O1 and O3. It seems that adding embeddings makes overfitting the machine learning model to the dataset. As a result, tests on other "kind of word understandability" and combined features are less successful compared to using standard features only for learning. This may also be due to the lack of systematicity in annotations of O2.

5.2 FrnnMUTE impact study

The FrnnMUTE embeddings were used separately and in combination with standard features and with FastText word embeddings for classifying medical words with the decision tree algorithm. To simplify the process of analyzing and comparing the results of this and the previous part, we aggregated the resulting F1 scores for combinations of a feature set and cross-validation scenario over all available users (Table 5). We observed that, in all cross-validation scenarios, our FrnnMUTE performs better when used separately by comparison with the FastText word embeddings used separately. FrnnMUTE provides the maximal F1 score (79.5) among user pairs versus the F1 score provided by the FastText word embeddings in user-in

Train user	*Test user*	*Standard features*				*FastText embeddings*				*Standard features + FastText embeddings*			
		A	P	R	F	A	P	R	F	A	P	R	F
$O1$	$O1$	**82.5**	77.2	**82.5**	79.8	72.5	67	72.5	69.3	82.4	**79**	82.4	**80.2**
$O2$	$O2$	**82**	78.9	**82**	80	73.5	69.9	73.5	71.3	81.9	**79.5**	81.9	**80.3**
$O3$	$O3$	85.5	81.2	85.5	83.2	74.9	70.4	74.9	72.3	**85.9**	**83**	**85.9**	**84.2**

Table 2: Experiments on user-in vocabulary-out cross-validation. The best score for a combination of quality measure and experiment is in bold.

Train user	*Test user*	*Standard features*				*FastText embeddings*				*Standard features + FastText embeddings*			
		A	P	R	F	A	P	R	F	A	P	R	F
O1	*O2*	81.7	78.6	81.7	80.1	74	70.3	74	71.2	**84.2**	**82**	**84.2**	**82.8**
O1	*O3*	85	81.2	85	83	75.4	70.7	75.4	72.6	**87.6**	**84.9**	**87.6**	**85.9**
O2	*O1*	82.2	77	82.2	79.1	72.8	67.3	72.8	69.6	**83.9**	**80.2**	**83.9**	**81.1**
O2	*O3*	85.4	81.1	85.4	83	75.3	71.1	75.3	73	**86.8**	**83.5**	**86.8**	**84.7**
O3	*O1*	82.8	77.4	82.8	79.7	72.7	67.1	72.7	69.4	**84.9**	**81.3**	**84.9**	**82.4**
O3	*O2*	82.2	79	82.2	80.2	74.1	70.4	74.1	71.6	**84.2**	**82.1**	**84.2**	**82.8**

Table 3: Experiments on user-out vocabulary-in cross-validation.

Train user	*Test user*	*Standard features*				*FastText embeddings*				*Standard features + FastText embeddings*			
		A	P	R	F	A	P	R	F	A	P	R	F
O1	*O2*	81.7	78.6	81.7	80.1	73.6	69.9	73.6	71.3	**81.8**	**79.8**	**81.8**	**80.6**
O1	*O3*	**85**	81.2	**85**	83	74.8	70.4	74.8	72.4	84.9	**82.2**	84.9	**83.4**
O2	*O1*	**82.2**	76.9	**82.2**	**79.1**	72.5	66.9	72.5	69.3	81.7	**77.5**	81.7	**79.1**
O2	*O3*	**85.3**	81	**85.3**	**83**	75.1	70.7	75.1	72.7	84.4	**81.3**	84.4	82.5
O3	*O2*	**82.7**	77.3	**82.7**	79.7	72.5	66.9	72.5	69.2	82.6	**78.9**	82.6	**80.2**
O3	*O3*	82.1	79	82.1	80.1	73.8	70.2	73.8	71.4	**82.2**	**80**	**82.2**	**80.7**

Table 4: Experiments on user-out vocabulary-out cross-validation.

	user-in vocabulary-out		user-out vocabulary-in		user-out vocabulary-out	
	$\mu \pm \sigma$	max	$\mu \pm \sigma$	max	$\mu \pm \sigma$	max
Standard features	77.7 ± 5.2	83.4	77.7 ± 4.9	84.4	77.6 ± 4.9	84.3
FT emb	67.9 ± 5.7	75.1	67.6 ± 5.3	75.3	67.3 ± 5.2	74.9
FrnnMUTE	75.1 ± 3.9	79.5	77.1 ± 3.9	82.4	74.5 ± 3.9	79.6
Standard features + FT emb	78.9 ± 5.1	85.2	79.5 ± 4.6	86.9	77.1 ± 4.6	84.6
Standard features + FrnnMUTE	80.0 ± 5.1	85.8	80.3 ± 4.3	87.0	78.6 ± 4.4	85.2
Standard features + FT emb + FrnnMUTE	79.9 ± 5.0	85.8	80.4 ± 4.3	87.4	78.1 ± 4.3	85.2

Table 5: Mean, standard deviation and maximum of F1 scores

vocabulary-out cross-validation (75.1). Similarly, the F1 score is higher on the user-out vocabulary-in experiment (82.4 versus 75.3), and in the user-out vocabulary-out experiment (79.6 versus 74.9). The FrnnMUTE results have the smallest dispersion (3.8-3.9) among all considered "solo" feature sets types (4.8-5.3) when aggregated by all available users. This means that FrnnMUTE are more robust in generalizing information from user to user and between different subsets of vocabulary. For the user-in vocabulary-out and the user-out vocabulary-out experiments the combination of standard features and FrnnMUTE in almost all cases shows the best performance among all feature sets. We can observe that the difference in F1 reaches 2.9 for some user pairs and that the maximum improvement achieved by combining standard features with FrnnMUTE over using standard features only hits 5.2 in F-measure. This testifies that FrnnMUTE helps standard linguistic and non-linguistic features to capture word understandability better than FastText embeddings. The fact that the combination of all three feature sets performs insignificantly better of even worse than standard features with only FrnnMUTE can be explained by the overfitting of the classification model in the first case because the resulting feature vector has the biggest dimensionality.

6 Conclusion

We tackle the prediction of understanding of French medical words by using FastText word embeddings as features. Yet, the embeddings solely as features are not enough for good word categorization. Whereas adding FastText word embeddings to standard features results in a substantial improvement of classification model performance when generalizing them to unknown users. We also proposed a novel type of embeddings trained on reference data from one annotator, and called them FrnnMUTE (French RNN Medical Understandability Text Embeddings). Compared with the case of using only standard features with and without FastText word embeddings, the combination of our FrnnMUTE with standard features substantially improves the performance of classification model. This indicates that FrnnMUTE capture better the specifics of medical words required for identifying their understandability by users, than FastText word embeddings. The FrnnMUTE embeddings and the categorization code are being made publicly available for scientific non-commercial purposes[8].

We have several directions for future work. Currently we use the existing word embeddings pretrained on Wikipedia and Web Crawl. We assume that training words embeddings on medical data may improve their impact on the results from categorization of medical terms. Another issue is that, after analysis of results of the application of FastText word embeddings in a categorization task, we assumed the existence of a robust nonlinear dependency between some subsets of standard features and subword-level components of FastText word embeddings. We plan to test this hypothesis in further research. Finally, while the annotations go forward, the annotators usually show *learning* progress in decoding the morphological structure of terms and their understanding. This progress is not taken into account in the current experiments, and is also the topic of our future research.

Acknowledgments

This work has been partly founded by the French ANR (grant number ANR-17-CE19-0016-01) as part of the project CLEAR (Communication, Literacy, Education, Accessibility, Readibility).

References

Joachim Bingel, Natalie Schluter, and Héctor Martínez Alonso. 2016. Coastalcph at semeval-2016 task 11: The importance of designing your neural networks right. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1028–1033. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

A Borst, A Gaudinat, C Boyer, and N Grabar. 2008. Lexically based distinction of readability levels of health documents. In *MIE 2008*. Poster.

J Chmielik and N Grabar. 2011. Détection de la spécialisation scientifique et technique des documents biomédicaux grâce aux informations morphologiques. *TAL*, 51(2):151–179.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

[8] https://github.com/hpylieva/FrnnMUTE

Roger A. Côté, D. J. Rothwell, J. L. Palotay, R. S. Beckett, and Louise Brochu. 1993. *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*. College of American Pathologists, Northfield.

Gunther Eysenbach. 2007. Poverty, human development, and the role of ehealth. *J Med Internet Res*, 9(4):34–4.

R Flesch. 1948. A new readability yardstick. *Journ Appl Psychol*, 23:221–233.

L Goeuriot, N Grabar, and B Daille. 2008. Characterization of scientific and popular science discourse in french, japanese and russian. In *LREC*.

N Grabar, 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*, pages 284–288.

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

Robert Gunning. 1973. *The technique of clear writing*. McGraw Hill, New York, NY.

JP Kincaid, RP Jr Fishburne, RL Rogers, and BS Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

D Kokkinakis and M Toporowska Gronostaj. 2006. Comparing lay and professional language in cardiovascular disorders corpora. In *WSEAS Transactions on BIOLOGY and BIOMEDICINE*, pages 429–437.

Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016. Ltg at semeval-2016 task 11: Complex word identification with classifier ensembles. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 996–1000. Association for Computational Linguistics.

Alexa T. McCray. 2005. Promoting health literacy. *Journal of the American Medical Informatics Association*, 12(2):152–163.

F Namer. 2000. FLEMM : un analyseur flexionnel du français à base de règles. *Traitement automatique des langues (TAL)*, 41(2):523–547.

Fiammetta Namer and Pierre Zweigenbaum. 2004. Acquiring meaning for French medical terminology: contribution of morphosemantics. In *Ann Symp Am Med Inform Assoc*.

M Poprat, 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*, pages 503–508, Maastricht.

Francesco Ronzano, Ahmed Abura'ed, Luis Espinosa Anke, and Horacio Saggion. 2016. Taln at semeval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1011–1016. Association for Computational Linguistics.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Int Conf on New Methods in Language Processing*, pages 44–49.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Luo Si and James P. Callan. 2001. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM '01)*, pages 574–576.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112.

Q Zeng-Treiler, H Kim, S Goryachev, A Keselman, L Slaugther, and CA Smith. 2007. Text characteristics of clinical reports and their implications for the readability of personal health records. In *MEDINFO*, pages 1117–1121, Brisbane, Australia.

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