Transfer-Based Learning-to-Rank Assessment of Medical Term Technicality

Dhouha Bouamor Leonardo Campillos-Llanos Anne-Laure Ligozat
Sophie Rosset Pierre Zweigenbaum
LIMSI, CNRS, Université Paris-Saclay, 91405 Orsay, France
firstname.lastname@limsi.fr

Abstract

While measuring the readability of texts has been a long-standing research topic, assessing the technicality of terms has only been addressed more recently and mostly for the English language. In this paper, we train a learning-to-rank model to determine a specialization degree for each term found in a given list. Since no training data for this task exist for French, we train our system with non-lexical features on English data, namely, the Consumer Health Vocabulary, then apply it to French. The features include the likelihood ratio of the term based on specialized and lay language models, and tests for containing morphologically complex words. The evaluation of this approach is conducted on 134 terms from the UMLS Metathesaurus and 868 terms from the Eugloss thesaurus. The Normalized Discounted Cumulative Gain obtained by our system is over 0.8 on both test sets. Besides, thanks to the learning-to-rank approach, adding morphological features to the language model features improves the results on the Eugloss thesaurus.

Keywords: Technicality of Medical Terms; Learning to rank; Terminology

1. Introduction

The difference between the language used by health care professionals and that used by patients is cited as a source of miscommunication (Elhadad and Sutaria, 2007) or difficulty when mining patient forums (Nikfarjam et al., 2015). For example, lay people tend to use idiomatic expressions such as *"mal de chien"(Fr) [*literally*, "dog pain" (En)]* to refer to *"douleur intense"(Fr) [severe pain, En]*.

Dictionaries which differentiate familiar terms from specialized terms are valuable for many applications of Natural Language Processing such as sentiment analysis (Ali et al., 2013) and paraphrase acquisition (Elhadad and Sutaria, 2007; Deléger and Zweigenbaum, 2009; van der Plas and Tiedemann, 2010). In the context of an e-learning system for training medical students, we are creating a virtual patient who must answer student questions in a natural way, hence using patient language (Campillos-Llanos et al., 2015; Campillos-Llanos et al., 2016). This is yet another application which requires a way to assess which term among a set of equivalent expressions would be more likely to be used by a patient.

Nevertheless, such resources remain scarce and thus far, only one, the Consumer Health Vocabulary (CHV, Keselman et al. (2007)), aims at estimating how familiar a term is. However, the CHV only covers the English language, and limited attempts have been made to cover other languages such as French and Portuguese. Tapi Nzali et al. (2015) mention the creation of a French CHV, but actually address a different problem: the identification of abbreviations, of spelling errors, and of lay terms associated to (but not necessarily equivalent to) specialized terms, such as *morphine* (morphin, which they consider as a lay term) – *douleur* (pain, which they consider as a specialized term). Tenorio and Torres Pisa (2015) describe plans to build a Brazilian Portuguese CHV, but only human assessment of the status of terms is foreseen.

In this paper, we describe a learning-to-rank approach that aims at assigning a specialization degree to each of the entries given in a list of synonym terms, based on statistical and morphological features. The newly introduced approach is based on learning-to-rank techniques where we combine statistical and linguistic features to determine how likely it is for a term to be used by lay or specialized speakers. Statistical features are extracted based on two languages models estimated on lay and specialized corpora. Linguistic features rely on morphological information including clues of being a derived word a neoclassical compound word. To the best of our knowledge, this is a first attempt to rank terms according to their specialization degree for the French language. Moreover, we adopt a transfer learning approach to build on data available for English to create a solution for French.

The remainder of the paper is organized as follows. Section 2 recalls some previous work. Section 3 presents the main contribution of this paper, which consists in ranking terms according to their specialization degree. Section 4 describes the experimental protocol we followed and Section 5 discusses the obtained results, then we conclude in Section 6.

2. Related work

This work can be associated to the research carried out in the field of text readability. Readability research studies how easily a text can be understood. Two types of readability measures are distinguished: classical and computational (François, 2011). Classical measures are essentially based on the number of characters and/or syllables in words, sentences or documents, and on linear regression models (Flesch, 1948). Computational measures might involve vector space models and a wide range of descriptors and their combinations (Zeng et al., 2005a; Wang, 2006; François and Fairon, 2013; Zeng-Treitler et al., 2007; Leroy et al., 2008). However, text readability formulas, which are mostly based on word length and sentence length, have been deemed inappropriate to measure the readability of medical texts: "counting words and syllables and consulting a grade-level word list are most likely not sufficient to determine how readable a text is" (McCray, 2005).

This has prompted researchers (Keselman et al., 2007) to design more appropriate measures for medical texts which

take into account "surface-level term familiarity, or recognition of the lexical form", and "concept-level term familiarity, or understanding of the underlying concept". They collected familiarity data from 41 subjects for training, and term and word frequencies in three different corpora were used as features. The algorithm assigned each consumer health term a predictive score ranging from 0 to 1.0, representing the likelihood that a term be familiar to the average consumer. Terms were classified into three familiarity categories based on their scores: "likely" (> 0.8), "somewhat likely" (0.8-0.5), and "not likely" to be familiar (scores $<$ 0.5).

Kandula and Zeng-Treitler (2008) created a gold standard set of 324 documents, based on expert ratings, to evaluate the readability of health texts, and observed that two commonly used readability formulae (FKGL and SMOG) did not have a very strong correlation with these ratings. Zeng et al. (2005b) identified 'consumer-friendly' names for 425 commonly used health concepts, based on corpus collection and manual review. Zeng et al. (2005a) predicted the average familiarity of 68 medical terms with an SVM classifier, using as features the term frequency in three health text corpora (specialized: MEDLINE; both specialized and lay: MedlinePlus; and lay: MedlinePlus log of user-submitted queries), the percentage of words from the Dale-Chall list of easy words, and the average word length. They obtained 0.196 mean absolute error (which we interpret as 80.4% accuracy).

Vydiswaran et al. (2014) used a Wikipedia-based measure to differentiate consumer terms from professional terms. The approach is based on comparing term frequency in MEDLINE, which indexes scientific papers produced by the professional community, to term frequency in MedHelp, a popular online health forum where content is mainly generated by lay people, and obtained 93.1% accuracy on a set of 58 pairs of consumer *vs.* professional terms.

In contrast to previous work,

1. Instead of comparing the frequencies of full terms in specialized and lay corpora as in (Zeng et al., 2005a) and (Vydiswaran et al., 2014), we compare the probabilities of these terms based on trigram *language models*. An advantage is that language models can estimate the likelihood of unseen terms, both through the probabilities of shorter n-grams and by using smoothing techniques.

2. To build on the existence of gold standard lists of familiarity-ranked terms, we adopt a learning-to-rank method.

3. Since these training data exist for the English language but not for our target language (French), we perform transfer learning from English training data to French test data.

3. Assessing the technicality of terms

The focus of this paper is to assess the technicality of a term by computing a specialization degree. In contrast to previous work on this topic, we propose to use a learning-to-rank based approach. Learning to rank is a relatively new field in which machine learning algorithms are used to learn a ranking function. It is useful for many NLP applications such as information retrieval (Liu, 2009), machine translation (Duh and Kirchhoff, 2008), or recommender systems (Lv et al., 2011).

In this work, the ranking function is learned from English data sets and is tested on French data. This is because, on the one hand, of the scarcity of annotated data with specialization degrees for French. On the other hand, assessing specialization degrees manually is difficult to carry out and is a time-consuming task. Data from the English CHV are used here as training data to estimate different models, and associated familiarity likelihood scores are used as labels.

Our approach combines statistical and linguistic features to assess term technicality. We present a statistical feature based on language modeling in Section 3.1 and linguistic features based on morphology in Section 3.2.

3.1. Statistical feature: likelihood ratio test based on two language models

The Likelihood ratio test between two language models is used as a statistical feature. The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than under the other. For this, two language models are estimated on two corpora: we consider a lay corpus and a specialized corpus as texts written in two different language levels. Forums are supposed to represent patient language, and scientific articles professional language. Even though some specialized terms may be used by patients in the forums, we expect them to occur more often in the specialized corpus. Hence the relative likelihood of a term for the language models estimated on the two corpora should give an indication of their degree of specialization.

We estimate a trigram language model on each of the corpora using SRILM[1]. We consider a trigram model, i.e., each word depends only on the previous two words; and we do not consider begin-of-sentence and end-of-sentence probabilities. This is consistent with the aim of the present work, where we would like to evaluate terms without considering the context in which they occur.

3.2. Linguistic features: morphology

Morphological features include the following:

- Being a *derived word* (binary feature): in specialized medical texts, authors tend to use relational adjectives where in most cases lay people use nouns (Deléger and Zweigenbaum, 2009). For instance, specialized authors tend to use the derived adjective *aortique [En aortic]* whereas lay people tend to use the base noun *aorte [En aorta]*.

- Containing *components* of neoclassical compound medical terms (binary feature): a term is more likely to be specialized if it contains morphological components typical of medical words. These components

[1] http://www.speech.sri.com/projects/srilm/

may occur in final position *("-graphy" in "radiography, nosography")*, initial position *("allo-" in "allograft, allopolyploid)* or root position *("mamm" in "mammary, xeromammography")*.

- *Term length* (numeric feature): specialized terms are longer than lay terms. Word length is used in readability formulae (Flesch, 1948) as well as in a classifier by (Zeng et al., 2005a).

Note that the features we rely on are not the derived words or components themselves, but being a derived word or containing a compound word component. This property of being non-lexical makes these features comparable and transferable across languages. In our experiments, we compute these features based on English resources for English terms then based on French resources for French terms.

4. Resources and experimental setup

Language models were trained on two pairs of two corpora, one pair per language:

English, specialized: a $30Mword$ sample of the PubMed Central (PMC[2]) corpus, composed of full-text English articles in the biomedical domain.

English, lay: the corpus provided by the CLEF eHealth 2015 information retrieval shared task[3] (50M words). This corpus consists of Web pages in French covering a broad range of health topics, targeted at the general public.

French, specialized: the CRTT corpus[4], a specialized corpus composed of full-text French scientific papers in the biomedical domain from Elsevier journals, containing about $25M$ words.

French, lay: discussions found in the French medical forum *atoute.org*[5] which covers several medical topics and is composed of 18M words.

To train a learning-to-rank model, we used the English Consumer Health Vocabulary and its familiarity scores as training set, the PMC and CLEF eHealth corpora to compute the specialized and lay language models, and the UMLS Specialist Lexicon resources to compute morphological features. The SVMMap toolkit (Yue et al., 2007), which is designed to optimize rankings for the Mean Average Precision (MAP) performance measure, was used to estimate ranking functions. It performs supervised learning using binary labeled training examples. We used a threshold to transform the familiarity scores of the English CHV into the required binary labels: -1 for lay terms and $+1$ for specialized terms. In our experiments, different values (0.1, 0.3, 0.5 and 0.7) were considered for this threshold.

[2] http://www.ncbi.nlm.nih.gov/pmc/

[3] https://sites.google.com/site/clefehealth2015/task-2

[4] http://crtt.univ-lyon2.fr/les-corpus-medicaux-du-crtt-613310.kjsp

[5] http://www.atoute.org/n/-forums-.html

To apply the model to French data, we computed specialized and lay language models on the CRTT and *atoute.org* corpora, and used morphological resources from the DeriF morphological analyzer (Namer and Zweigenbaum, 2004) and from the UMLF lexicon (Zweigenbaum et al., 2005).
We evaluated the model on two French data sets: EUGLOSS and UMLS. EUGLOSS[6] is a multilingual glossary of technical and popular medical terms. Our test collection is composed of 868 lay terms and their corresponding technical medical terms. Table 1 lists sample entries, for which we show English translations for convenience (these translations are not used in our system).

lay	specialized
trouble de la vue *vision disorder*	diplopie *diplopia*
trouble de la digestion *impaired digestion*	dyspepsie *dyspepsia*
trouble du système nerveux *nervous system disorder*	athétose *athetosis*
sous la peau *under the skin*	sous-cutané *subcutaneous*

Table 1: French lay and specialized terms in EUGLOSS. English translations are provided for convenience and do not necessarily belong to EUGLOSS.

The French part of the UMLS Metathesaurus contains both lay and specialized terms, but no information about their technicality is given. Therefore, we manually ranked 134 UMLS terms associated with 32 CUIs (concept identifiers) from lay to specialized. An example is presented in Table 2.

infarctus myocardique	1
infarctus du myocarde	2
crise coronaire	3
im	4
crise cardiaque	5

Table 2: Manually ranked terms from specialized to lay, having a CUI of C0027051 (Myocardial Infarction). 1 is most specialized, 5 is least specialized.

5. Results and discussion

The evaluation of our approach was performed using the Normalized Discounted Cumulative Gain (NDCG). NDCG is a measure of ranking quality used in information retrieval. The performance of our learning-to-rank approach was compared to using only the likelihood ratio test to rank terms, as presented in Section 3.1, denoted as LLR_{Test}. This allows us to study the effectiveness of the addition of linguistic features through the learning-to-rank method. Table 3 describes the obtained results.
The first notable observation is that the learning-to-rank approach outperforms LLR_{Test} on the EUGLOSS test set for almost all configurations. This shows the positive contribution of the information brought by linguistic features. The

[6] http://users.ugent.be/~rvdstich/eugloss/

	EUGLOSS	UMLS
LLR_{Test}	0.79	**0.89**
$LTR_{0.1}$	0.80	0.69
$LTR_{0.3}$	**0.83**	0.66
$LTR_{0.5}$	0.82	0.66
$LTR_{0.7}$	0.79	0.64

Table 3: NDCG values on the EUGLOSS and UMLS test sets. LLR_{Test}: baseline system. LTR_i: learning-to-rank method. i: threshold above which a term is considered as lay to train LTR.

highest NDCG is obtained by $LTR_{0.3}$. Considering as lay terms CHV terms having a familiarity score greater than 0.3 gives the best results here.

However, different results were obtained for the UMLS test set. Considering only likelihood ratio test values for the ranking task proved sufficient: a high score of 0.89 was obtained. Let us recall though that the UMLS test set was annotated manually by a lay person: this might explain the difference between the results obtained on the two test sets. In contrast, the learning-to-rank approach was effective for the EUGLOSS test set, which was ranked with the help of domain specialists. Moreover, the size of the UMLS test set is smaller than the EUGLOSS test set. This also makes the results on the UMLS test set less reliable.

6. Conclusion

We presented in this paper an approach which assesses the technicality of terms. The proposed approach is based on learning-to-rank, where statistical and linguistic features are combined to determine how likely a term is to be associated with a lay or specialized audience. The obtained results show that, when applied to EUGLOSS, which was created by health care professionals, linguistic features help statistical features in the ranking process.

The quality of the obtained ranking is high, despite the transfer from English to French, and was tested on a much larger test set than (Zeng et al., 2005a) and (Vydiswaran et al., 2014).

The obtained results are thus very encouraging, but experiments can still be improved in a number of ways. First, we plan to expand the UMLS test set and ask more than one annotator to rank terms according to their specialization degree, additionally computing an inter-annotator agreement. Then, we plan to apply our approach to create technicality rankings for the whole French part of the UMLS, thus providing a valuable resource to the community.

7. Acknowledgements

This work was partly funded by BPI through the FUI Project PatientGenesys (F1310002-P).

8. References

Ali, T., Schramm, D., Sokolova, M., and Inkpen, D. (2013). Can I hear you? sentiment analysis on medical forums. In *IJCNLP*, pages 667–673.

Campillos-Llanos, L., Bouamor, D., Bilinsky, E., Ligozat, A.-L., Zweigenbaum, P., and Rosset, S. (2015). Description of the PatientGenesys dialogue system. In *Proc. of 16th SIGDIAL*, pages 438–440.

Campillos-Llanos, L., Bouamor, D., Zweigenbaum, P., and Rosset, S. (2016). Managing linguistic and terminological variation in a medical dialogue system. In *LREC 2016, Tenth International Conference on Language Resources and Evaluation*, Portoroz, Slovenia. ELRA. This volume.

Deléger, L. and Zweigenbaum, P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proc BUCC Workshop*, pages 2–10. ACL.

Duh, K. and Kirchhoff, K. (2008). Learning to rank with partially-labeled data. In *SIGIR*, pages 251–258. ACM.

Elhadad, N. and Sutaria, K. (2007). Mining a lexicon of technical terms and lay equivalents. In *Proc BioNLP Workshop*, pages 49–56. ACL.

Flesch, R. (1948). A new readability yardstick. In *Journal of Applied Psychology*, pages 221–233.

François, T. and Fairon, C. (2013). Les apports du TAL à la lisibilité du français langue étrangère. *TAL*, 54(1):171–202.

François, T. (2011). *Les apports du traitements automatique du langage à la lisibilité du français langue étrangère.* Ph.D. thesis, Université Catholique de Louvain, Louvain, Belgium.

Kandula, S. and Zeng-Treitler, Q. (2008). Creating a gold standard for the readability measurement of health texts. In *AMIA Annu Symp Proc*, pages 353–357.

Keselman, A., Tse, T., Crowell, J., Browne, A., Ngo, L., and Zeng, Q. (2007). Assessing consumer health vocabulary familiarity: an exploratory study. *J Med Internet Res*, 9(1):e5.

Leroy, G., Helmreich, S., Cowie, J., Miller, T., and Zheng, W. (2008). Evaluating online health information: Beyond readability formulas. In *AMIA Annual Symposium Proceedings*.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, mar.

Lv, Y., Moon, T., Kolari, P., Zheng, Z., Wang, X., and Chang, Y. (2011). Learning to model relatedness for news recommendation. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 57–66, New York, NY, USA. ACM.

McCray, A. T. (2005). Promoting health literacy. *J Am Med Inform Assoc*, 12(2):152–163, Mar-Apr.

Namer, F. and Zweigenbaum, P. (2004). Acquiring meaning for french medical terminology: contribution of morphosemantics. *Eleventh MEDINFO International Conference*, pages 535–539.

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., and Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc*, 22(3):671–681.

Tapi Nzali, M. D., Bringay, S., Lavergne, C., Opitz, T., Azé, J., and Mollevi, C. (2015). Construction d'un vo-

cabulaire patient/médecin dédié au cancer du sein à partir des médias sociaux. In *IC2015*, IC2015, Rennes, France, June.

Tenorio, J. M. and Torres Pisa, I. (2015). Consumer Health Vocabulary: A proposal for a Brazilian Portuguese language. *Stud Health Technol Inform*, 216:1089.

van der Plas, L. and Tiedemann, J. (2010). Finding medical term variations using parallel corpora and distributional similarity. In *Proceedings of the 6th Workshop on Ontologies and Lexical Resources* , pages 28–37, Beijing, China, August. Coling 2010 Organizing Committee.

Vydiswaran, V. G. V., Mei, Q., Hanauer, D. A., and Zheng, K. (2014). Mining consumer health vocabulary from community-generated text. In *AMIA Annu Symp Proc*, volume 2014, pages 1150–1159.

Wang, Y. (2006). Automatic recognition of text difficulty from consumers health information. In *CBMS*, pages 131–136. IEEE Computer Society.

Yue, Y., Finley, T., Radlinski, F., and Joachims, T. (2007). A support vector method for optimizing average precision. In *SIGIR*, pages 271–278. ACM.

Zeng, Q., Kim, E., Crowell, J., and Tse, T. (2005a). A text corpora-based estimation of the familiarity of health terminology. In *ISBMDA*, pages 184–192, Aveiro, Portugal.

Zeng, Q. T., Tse, T., Crowell, J., Divita, G., Roth, L., and Browne, A. C. (2005b). Identifying consumer-friendly display (CFD) names for health concepts. In *AMIA Annu Symp Proc*, pages 859–863.

Zeng-Treitler, Q., Kim, H., Goryachev, S., Keselman, A., Slaughter, L., and Smith, C. A. (2007). Text characteristics of clinical reports and their implications for the readability of personal health records. *Stud Health Technol Inform*, 129(Pt 2):1117–1121.

Zweigenbaum, P., Baud, R. H., Burgun, A., Namer, F., Jarrousse, E., Grabar, N., Ruch, P., Le Duff, F., Forget, J.-F., Douyère, M., and Darmoni, S. (2005). A unified medical lexicon for French. *International Journal of Medical Informatics*, 74(2–4):119–124.