Application of cross-language criteria for the automatic distinction of expert and non expert online health documents

Natalia Grabar[1,2], Sonia Krivine[3]

[1] INSERM, UMR_S 729, Eq. 20, Paris, F-75006 France
[2] Health on the Net Foundation, SIM/HUG, Geneva, Switzerland
[3] FircoSoft, 37 rue de Lyon, 75012 Paris, France
natalia.grabar@spim.jussieu.fr, sonia.krivine@free.fr

Abstract. Distinction between expert and non expert documents is an important issue in the medical area, for instance in the context of information retrieval. In our work we address this issue through stylistic corpus analysis and application of machine learning algorithms. Our hypothesis is that this distinction can be observed on the basis of a little number of criteria and that such criteria can be language and domain independent. The used criteria have been acquired in source corpus (Russian) and then tested on source and target (French) corpora. The method shows up to 90% precision and 93% recall, and 85% precision and 74% recall in source and target corpora.

1 Introduction

Medical information searchable online presents various technical and scientific content but this situation is not clear for non expert users. As a matter of fact, when reading documents with high technical content non expert users can have some understanding problems, because they are anxious, pressed or unfamiliar with the health topic. This situation can have a direct impact on users' well-being, their healthcare or communication with medical professionals. For this reason, search engines should distinguish documents according to whether they are written for medical experts or non expert users. Distinction between expert and non expert documents is closely related to the health literacy [1], and the causal effect it can have on healthcare [2]. For the definition of the readability level, several formulae have been proposed (*i.e.*, Flesch [3], Fog [4]), which rely on criteria like average length of words, sentences and number of difficult words. Distinction between expert and non expert documents can also be addressed through algorithms proposed by the area of text categorisation and applied to various features: Decision Tree and Naive Baayes applied to manually weighted MeSH terms [5]; TextCat[4] tool applied to *n-grams* of characters [6]; SVM applied to a combination of various features [7].

[4] *www.let.rug.nl/∼vannoord/TextCat*

In our work, we aim at applying machine learning algorithms to corpora which gather documents from different languages and domains. To ease this process, we propose to use a little set of features, which would be easy to define and to apply to a new language or domain. Aimed features should be shared by different languages and domains. Assuming that documents represent the context of their creation and usage through both their content and style, we propose to set features at the stylistic level. Features are thus defined on the basis of the source corpus and then applied to the target corpus. Languages and domains of these two corpora are different. The cross-domain and especially cross-language aspect of features seems to be a new issue in the text categorisation area.

2 Material and Method

Working languages are Russian (source language) and French (target language). Corpora are collected online: through general engines in Russian and the specific medical search engine CISMeF in French. In Russian, the keywords used are related to *diabetis and diet*, and the distinction between expert and non expert dociments is performed manually. The French search engine already proposes this distinction and we exploit it in our work. We used keyword *pneumologie* (*pneumology*) when querying CISMeF. Table 1 indicates size and composition of corpora in both studied languages. The French corpus contains more documents, which is certainly due to the current Internet situation. Moreover, we can observe difference between sizes of expert and non expert corpora: the non expert corpus is bigger in Russian, while the expert corpus is bigger in French.

Table 1. Expert and non expert corpora in Russian and French languages.

	Russian		French	
	nbDoc	occ	nbDoc	occ
Expert documents	35	116'000	186	371'045
Non expert documents	133	190'000	80	87'177
Total	168	306'000	266	458'222

The objective of our work is to develop tools for categorising health documents according to whether they are expert or non expert oriented. We use several machine learning algorithms (`Naive Bayes`, `J48`, `RandomForest`, `OneR` and `KStar`) within Weka[5] tool in order to compare their performances and to check the consistency of the feature set. The main challenge of the method relies on the universality of the proposed features defined on the basis of source language (Russian) and domain (diabetis) and then applied to target language (French) and domain (pneumonology).

[5] Weka (*Waikato Environment for knowledge analysis*), developped at University Waikato, New-Zeland, is freely available on *www.cs.waikato.ac.nz/∼ml/index.html*

Table 2. Evaluation of algorithms on source and target corpora

Method	Expert		Non expert	
	Prec.	Recall	Prec.	Recall
NaveBayes	43	83	94	72
J48	**83**	**42**	86	98
RandomForest	**83**	**42**	86	98
OneR	43	25	82	91
KStar	70	58	**90**	**93**

Method	Expert		Non expert	
	Prec.	Recall	Prec.	Recall
NaveBayes	93	36	31	91
J48	81	83	43	41
RandomForest	**87**	**81**	**52**	**64**
OneR	83	87	53	45
KStar	85	74	42	59

Stylistic features have emerged from a previous contrastive study of expert and non expert corpora in Russian [8] realised with lexicometric tools. For the current work, we selected a set of 14 features related to the document structure, and marks of persons, punctuation and uncertainty. Learning and test corpora are composed of respectively 66% and 33% of the whole corpora collected. Evaluation is done on independent corpus through classical measures: precision, recall, F-measure and error rate.

3 Results and Discussion

Results obtained on the Russian corpus are presented in the left part of table 2. For each method (first column), we indicate figures of precision and recall. `KStar` shows the best results with *non expert* documents: 90% precision and 93% recall, and nearly the best results for the *scientific* category: 70% precision and 58% recall. `J48` and `RandomForest`, both using decision trees, present identical results for two studied categories: 83% precision and 42% recall for *scientific* documents and 86% precision and 98% recall for *non expert* documents. From the point of view of precision, these two algorithms are suitable for the categorisation of documents as *scientific*. The right part of table 2 indicates evaluation results of the same algorithms applied to the French corpus (175 documents for learning and 91 for test). `RandomForest` has generated the most competitive results for both categories (*expert* and *non expert*). Surprisingly, `OneR`, based on the selection of only one rule, produced results which are close to those of `RandomForest`. As general remark, scientific documents are better categorised in French and non expert documents in Russian, which is certainly due to a larger size of corresponding data in each language. Low performances of `NaveBayes` in both languages seem to indicate that the Bayes model, and specifically its underlying hypothesis on independance of criteria, is too naive for the task of classification of documents as expert and non expert oriented. Whereas, we assume that stylistic and discourse criteria equally participate in the encoding of stylistic specificities of medical documents [9, 10].

Language model. We could analyse two language models, generated by `OneR` and `J48` algorithms. `OneR` selects one (best) rule in each corpus. In our experiment, this algorithm selected hypertext link <a> tag in Russian and 2^{nd}

plural pronoun in French. These features allow to produce nearly the best results in the target corpus (French), while in Russian this algorithm is the less competitive. The model produced by `J48` in Russian selects hypertext link <a> tags together with 1^{st} singular pronoun я (*I*), italic characters (tag <i>), lists (<ol>) and table (<table>) tags. On French corpora, `J48` selects the following five criteria: 2^{nd} plural pronoun, <table> tag, 2^{nd} singular pronoun, <ol> tag and exclamation mark. `J48` is one of the most suitable algorithms in Russian but it shows average performances in French. Surprisingly, a majority of the most relevant criteria are related to the `HTML` tagging of documents but not to linguistic information. This observation seems to indicate that categorisation of web documents should be based also on non textual criteria. According to the theory of genres [11], this observation emphasizes the importance of the layout of documents, their typography and intertextuality.

Analysis of errors common to various classifiers. Within Russian corpus, six documents are wrongly categorised by several algorithms. Their analysis indicates that these documents are ambiguous as for their categorisation, both manual and automatic, and that discourse distinction between expert and non expert document is set on a continuum axis. Thus, there is no dichotomy between these two categories and borderline documents are difficult to categorise.

Suitability of proposed features. The proposed reduced set of features contains 14 criteria related to the document structure, marks of persons, punctuation and uncertainty. The obtained results seem to indicate that these stylistic features are suitable for the categorisation of documents according to their discourse (*expert* and *non expert*). Indeed, their application to the target corpus shows promising performances, although the target corpus is composed of documents in a different language and describing different medical topics. Moreover, these features are easy to adapt to a new language. But their application to other corpora has to be verified. One of their limitations is that several of these features remain specific to HTML documents.

4 Conclusion and Perspectives

We have presented an experiment on automatic distinction of expert and non expert Web documents. For this, learning algorithms and a set of 14 stylistic criteria have been used. Criteria have been acquired on a source corpus (Russian language, diabetis related topic) and then applied to target (French language, pneumonology related topic) and source corpora. Evaluation results show that decision tree algorithms `J48` and `RandomForest` are the most suitable for the categorisation of documents as expert and non expert. They generate the best results in the target corpus and for the *expert* category in the source corpus. As we have noticed, results depend on the size of learning corpora. It would be interesting to apply the system to a larger collection of documents and to confirm the stability of the acquired language models. But we can consider these results as promising, especially as documents are extracted from various websites and learning and test steps are performed on independent datasets.

The obtained results seem to indicate that the proposed stylistic features are suitable for the categorisation of documents according to their discourse: their application to the target corpus shows promising performances, although the target corpus is composed of documents in different languages and describing different medical topics. Nevertheless, it could be interesting to apply other criteria, for instance argumentation structures [12], for the distinction between expert and non expert documents.

We assume that categorisation results can be more precise. For instance, within the category *scientific* we can distinguish scientific articles and didactical material; and within *non expert* category we can distinguish cook recipes, articles for large audience and food recommandations. In French, we built an intermediate category composed of documents written for medical students: *courses*, *teaching material*. It could be interesting to categorise this material through the proposed language model. It could be interesting to apply our method to other medical areas and other types of documents (clinical), and to compare it with results produced by other approaches.

References

1. McCray, A.: Promoting health literacy. Journal of American Medical Informatics Association **12** (2005) 152–163
2. AMA: Health literacy: report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. JAMA **281**(6) (1999) 552–7
3. Flesch, R.: A new readability yardstick. Journal of Applied Psychology **23** (1948) 221–233
4. Gunning, R.: The art of clear writing. McGraw Hill, New York, NY (1973)
5. Zheng, W., Milios, E., Watters, C.: Filtering for medical news items using a machine learning approach. In: AMIA. (2002) 949–53
6. Poprat, M., Markó, K., Hahn, U.: A language classifier that automatically divides medical documents for experts and health care consumers. In: MIE 2006 - Proceedings of the XX International Congress of the European Federation for Medical Informatics, Maastricht (2006) 503–508
7. Wang, Y.: Automatic recognition of text difficulty from consumers health information. In IEEE, ed.: Computer-Based Medical Systems. (2006)
8. Krivine, S., Tomimitsu, M., Grabar, N., Slodzian, M.: Relever des critères pour la distinction automatique entre les documents médicaux scientifiques et vulgarisés en russe et en japonais. In: TALN. (2006)
9. Benveniste, E.: La nature des pronoms. Problmes de linguistique gnrale **1** (1966) 251–257
10. Malrieu, D., Rastier, F.: Genres et variations morphosyntaxiques. in Traitement automatique des langues, vol. 42, p. 548-577 (2001)
11. Genette, G.: Théorie des genres. Seuil, Paris (1986)
12. Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbhler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D., Lovis, C., Veuthey, A.: Using argumentation to extract key sentences from biomedical abstracts. Int J Med Inform (2006)