INCLUDING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING INTO INFORMATION RETRIEVAL

Piotr Malak and Artur Ogurek

Institute of Information Science and Book Studies, University of Wrocław, Poland

ABSTRACT

In current paper we discuss the results of preliminary, but promising, research on including some Natural Language Processing (NLP) and Machine Learning (ML) approaches into Information Retrieval. Classical IR uses indexing and term weighting in order to increase pertinence of answers given to users queries. Such approach allows for matching the meaning, i.e. matching all keywords of the same or very similar meaning as expressed in user query. For most cases this approach is sufficient enough to fulfil user information needs.

However indexing and retrieving information over professional language texts brings new challenges as well as new possibilities. One of challenges is different grammar, causing the need of adjusting NLP tools for a given professiolect. One of the possibilities is detecting the context of occurrence of indexed term in the text.

In our research we made an attempt to answer the question whether Natural Language Processing approach combined with supervised Machine Learning is capable of detecting contextual features of professional language texts.

KEYWORDS

Enhanced Information Retrieval, Contextual IR, NLP, Machine Learning,

1. INTRODUCTION

Internet become the medium of diversified languages on different levels of communication process. One can find there texts given not only in different native languages, but also in different types of a given language. From linguistic perspective we may distinguish rich sets of texts in common language, in a dialect or a slang language, and professional texts, given in professiolect (a specific language of a given profession). All different types of one native language provide different morphosyntactical and semantic characteristics and features. In Natural Language Processing (NLP) approach they can be considered as a different languages, which shares the vocabulary with common language but have own grammatical rules and own sentence building rules. Professiolect texts base on common language vocabulary and grammar but use them in a characteristic way. The differences are frequencies of particular words usage, style of expression, voice, etc. As such they need adjusted processing rules.

Specific discipline language is also a challenge for available NLP tools. Typically NLP tools are trained on the base of a general language. As examples we may call Penn Treebank news texts

Natarajan Meghanathan et al. (Eds) : NLP, ARIA, JSE, DMS, ITCS - 2019
pp. 13-18, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.91202

corpora on Wall Street Journal (WSJ) or Universal Dependencies (UD) taggers. Effectivity of Part of Speech (PoS) tagging for WSJ reaches 97% (while for UD equals 94%) for general English texts, while for unknown words the effectivity may decrease for 10%. [Yu, Falenska & Vu, 2017].

Examples of professiolect are professional language of IT, microbiology or law e.g.. Any discipline is capable of creating professional language with own vocabulary set and grammatical rules. Our research focused on Polish judicial texts and the role of a subject in described situations. The role of a subject in judicial language is a function of context of meaning.

1.1. Previous research

Applying NLP for Information Retrieval (IR) purposes has a long tradition. Since beginning of automatic full texts indexing the language features of texts have been considered. Statistical approach is widely used for documents retrieval [Jurafsky, Martin, 1999]. In [Manning, Schütze, 1999] we may read about applications of statistical approaches for language analyses. Processing of Natural Language texts leads e.g. to automatic text categorisation (unsupervised detecting of similar texts on the basis of their language features) and to classification (supervised, trained process) [Jackson, Mouliner, 2002]. Development of NLP tools and their application in IR is described e.g. in [Brants, 2003].

Also using Machine Learning (ML) is well known approach in documents and texts retrieval. ML also plays important role in NLP. Using ML for text categorisation describes e.g. [Sebastiani, 2002]. Evaluation of performance of different classification approaches is presented e.g. in [Basili, Moschitti & Pazienza, 2001]. More general description of ML techniques is available in [Si, Jin, 2011].

However, using NLP and ML for IR was provided in general purposes. Typical examples are finding similar documents on the basis of their vector representations or automatic grouping of set of documents into subsets on the basis of subject or topic detection. In current paper we describe using well known approaches for quite new purpose – detecting contextual surrounding of chosen (named) entities. Preliminary research was made on judicial Polish language texts. Judicial language is an example of a professiolect. Entities (named and general) may appear in judicial texts in different contexts. The context depends on the role of described subject in the trial, which may be: claimant, defendant or participant (see next sections). This feature of the language may be used in order to provide more specific retrieval of information, with respect to users' needs. Distinction of documents or texts according to the role of entity may be achieved by combination of categorisation tasks and supervised machine learning for role recognition.

1.1.1 Judicial Polish language as different from common language

In 1948 [Wróblewski, 1948] noticed the differences between common Polish language and language being used in Polish law. Starting from those differences he proposed distinction of the judicial language from common lalnguage. Also [Malinowski, 2006; Choduń, 2006] discussed different features of law language in comparison to common language, on the example of Polish language. Malinowski, comparing and analysing law professiolect suggested additional distinction between judicial and law languages, by their different functional features.

In general judicial language uses only part of the whole vocabulary and characteristic grammar constructions and they provide different frequencies of the usage of the vocabulary. Judicial language in general is characterised by high precision of expressions, depersonalisation of expressions, high pertinence between a concept and its name, high ratio of noun terms in expression [Malinowski, 2006]. The language of judicial decisions differs from common

language in the means of vocabulary and grammar [Petzel, 2017], thus it require adjusted IR methods.

However IR does not offer methods for distinction between different grammatical forms – all words are typically transformed into basic stems (stemming) or into basic grammatical form (lemmatization). In order to increase pertinence of information retrieval within judicial documents texts Machine Learning (ML) methods may be applied.

1.2. Position of a subject in judicial texts

Despite of native language, all judicial texts have common features considering the position of a subject in described issue. A subject, recognized as named entity, may be a person or an institution involved in legal issue. Given entity may appear, in judicial text, in one of the following three positions:

- claimant – a person or an institution making a legal complaint against other person or institution,
- defendant – a person or an institution law case being accused of having done something illegal,
- participant – any entity occurring in the law case not as claimant nor defendant.

Depending on the position of a given entity in a law case, expressed in user query as a searching term, one could expect at least splitting the answer set for three disjoint subsets. Distinction between different position of a given entity, expressed by textual context of occurrence of entity in a document, is out of the reach of current Information Retrieval approaches.

Automatic classification may deliver sufficient solution. Our assumption was to train, in a supervised process, the classificator in order to enhance quality of information retrieval system answer to user queries. The deliverables of such enhancement of IR process may be set as follow:

- provide context distinction of matched documents on the basis of the role of query term (law case entity),
- provide such additional enhancement of IR process over professiolect documents in real time, without any delay in system answer time.

2. DESCRIPTION OF THE RESEARCH

In our research we prepared training-testing set of judicial documents of Polish law cases. As a testing entity we set *library*.

2.1. Corpora

From all retrieved documents in around 3% a library was in the role of a claimant – it was acting as legal institution. In around 14% of cases a library, as an institution was defendant. And remaining 83% was documents describing cases of participation of the subject in legal cases. The participant, however, appeared to be difficult to determine even manually. As a legal entity a Library may be subunit of other institutions as Universities, schools, prisons or local administrative units. In most of retrieved cases *library* was involved directly, but claimant or defendant was its superior unit. As a subject, *library* could be indirectly involved as well as act as situation neighbourhood – in many of described cases library was mentioned as a prison facility open for prisoners. Thus we decided to label such documents as *occurring* instead of juridical proper *participant*.

2.2. Setting of experiment

For labelling documents we applied CLEF-based approach, as described in [Malak, 2013]. Annotators received short instructions how to evaluate the role of a given subject in a document. Each document was then manually categorized to one of three following classes:

1. claimant,
2. defendant,
3. occurring.

Next all classical text pre-processing operations was conducted over training set of documents. All documents were subject to lemmatization and Part-of-Speech detection. For this stage NLP tools provided by Clarin-Pl were used to unify the grammatical forms of words in judicial decisions texts (morpho-syntactic tager WCRFT2 which joins *Conditional Random Fields* (CRF) and *tiered tagging* of plain text), programs that recognize the features characteristics of the text (Clarin-Pl: Fextor2) [Walkowiak, 2016] (most of these programs has been produced within the framework of Clarin-PL consortium).

Next feature vectors were generated using a bag-of-word method [Boulis and Ostendorf, 2002] and composed of frequencies of terms (a grammatical form of a word used in texts) in a document. Classification and validation was made according to [Walkowiak, Malak, 2018]. For classification runs we used the stratified k-fold cross-validation (with 4 folds) [Hastie et al., 2013]. The following algorithms were used in our experiments:

- Linear SVM with elastic net penalty learned by stochastic gradient descent (SVM_en_SGD) [Tsuruoka, Tsujii and Ananiadou, 2009],
- Multilayer Perceptron (MLP) [Hastie et al., 2013],
- Logistic Regression [Hastie et al., 2013],
- Decision Tree [Hastie et al., 2013],

2.3. Results

For the first stage of the research we achieved accuracy of classification ratio between 65% and 78%. For detecting *claimant* role it was 75% pertinence of retrieved, categorized documents, and for *defendant* role around 73%. As *occurring* role included cases where chosen subject (*library*) could be indirectly involved as well as act as situation neighbourhood, the pertinence ration of categorisation was lower, and achieved 69% in average.

As for classification algorithms used for this stage of research, the most promising efficiency offered linear SVM (SVM_en_SGD) as observed also in [Walkowiak, Malak, 2018].

3. CONCLUSIONS

The initial classification runs proves using NLP and classification into Information Retrieval may increase the pertinence of retrieval of professiolect texts. However, we still need to provide more analyses for the classification process. In our initial experiment we skipped reduction and weighting techniques. In further research we plan to apply *tf-idf* implementation, where the *tf* is normalized by max, and *ltu* weighting, although those methods are already used in IR indexing processes. We also want to examine POS filtering for nouns and verbs only, as [Savoy, 2006] proves the efficiency and accuracy of light stemming for French, Portuguese, German and Hungarian languages in context of Information Retrieval (IR). And, finally using a stoplist on

pre-processing stage for juridical texts classification should be also tested for its influence for accuracy of classification.

REFERENCES

[1] Basili, R., Moschitti, A., Pazienza, M. T., (2001). NLP-driven IR: Evaluating Performances over a Text Classification task. 1286-1294.

[2] Boulis, C., Ostendorf, M., (2002) Text classifcation by augmenting the bag-of-words representation with redundancy-compensated bigrams, In: Linguist Computing 17 (3): 267-287.

[3] Brants, T., (2003). Natural Language Processing in Information Retrieval, [on-line: https://www.researchgate.net/publication/220778280_Natural_Language_Processing_in_Information_Retrieval]

[4] Choduń A., (2006) Leksyka tekstów aktów prawnych, "Ruch Prawniczy, Ekonomiczny i Socjologiczny" 2006, z. 4.

[5] Hastie, T.J., Tibshirani, R.J., Friedman, J.H., (2013). The elements of statistical learning: data mining, inference, and prediction. Springer series in statistics, Springer, New York.

[6] Jackson, Peter; Moulinier, Isabelle (2002), Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization, Amsterdam/Philadelphia: John Benjamins Publishing Company

[7] Jurafsky, D., Martin, J. H. (1999), Speech and language processing. An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, New Jersey: Prentice Hall.

[8] Manning, Ch. D.; Schütze, H., (1999), Foundations of Statistical Natural Language Processing, Cambridge

[9] Malak P., (2013) Information searching over Cultural Heritage objects, and press news, 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, December 7-9, 2013, Poznań, Poland, ISBN 978-83-932640-4-9, 434-438.

[10] Malak P., (2018) Text Preprocessing: A Tool of Information Visualization and Digital Humanities, W: Information Visualization Techniques in the Social Sciences and Humanities, Osińska V., Osiński G. (red.), IGI Global 2018, s. 86-104.

[11] Malinowski A., Polski język prawny. Wybrane zagadnienia, Lexis-Nexis: Warszawa 2006

[12] Petzel J., (2017) Systemy wyszukiwania informacji prawnej, Wolters Kluwer.

[13] Savoy, J. (2006), Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages. Proceedings ACM-SAC, 1031-1035. The ACM Press.

[14] Sebastiani F., (2002), Machine learning in automated text categorization, ACM Computing Surveys (CSUR), v. 34 issue 1., 03.2002, pp. 1-47.

[15] Si, L., Jin, R., (2011), Machine learning for information retrieval. 1293-1294. 10.1145/2009916.2010167.

[16] Tsuruoka, Y., Tsujii, J., Ananiadou, S., (2009) Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In: ACL '09, 477–485.

[17] Walkowiak T. (2016) Asynchronous System for Clustering and Classifications of Texts in Polish. In: Zamojski W., Mazurkiewicz J., Sugier J., Walkowiak T., Kacprzyk J. (eds) Dependability Engineering and Complex Systems. Advances in Intelligent Systems and Computing, vol 470. Springer, Cham.

[18] Walkowiak T., Malak P., (2018) Polish Texts Topic Classification Evaluation, W: ICAART 2018 10th International Conference on Agents and Artificial Intelligence. Proceedings, vol. 2, Funchal, Madeira, Portugal 2018, s. 515-522.

[19] Wróblewski B., (1948) Język prawny i prawniczy. Kraków : Polska Akademia Umiejętności. 1948

[20] Yu X., Faleńska A. & Vu N. T. (2017) A general-purpose tagger with convolutional neural networks, in Proceedings of the First Workshop on Subword and Character Level Models in NLP, s. 124-129.