Polish Texts Topic Classification Evaluation

Tomasz Walkowiak[1] and Piotr Malak[2]
[1]*Faculty of Electronics, Wroclaw University of Science and Technology, Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland*
[2]*Institute of Information and Book Science, University of Wroclaw, pl Uniwersytecki 9-13. 50-140 Wroclaw, Poland*

Keywords: NLP, Polish, Text Classification, Feature Selection, Weighting Schema, Supervised Machine Learning.

Abstract: The paper presents preparation, lead and results of evaluation of efficiency of text classification (TC) methods for Polish. The subject language is of complex morphology, it belongs to flexional languages. Thus there is a strong need of making proper text preprocessing in order to guarantee reliable TC. Basing on authors' practical experience from former TC, IR and general NLP experiments set of preprocessing rules was applied. Also feature-documents matrix was designed with respect to the most promising feature selected. About 216 experiments on exemplar corpus in subject (topic) classification task, with different preprocessing, weighting, filtering (for dimensions reduction) schemes and classifiers was conducted. Results shows there is not substantial increase of accuracy when using most of classical pre-processing steps in case of corpus of large size (at least 1000 exemplars per class). The highest impact authors were able to obtain concerned the system costs of TC processes, not the TC accuracy.

1 INTRODUCTION

In this paper, we describe the main features, results and conclusions of a set of experiments in automatic subject (topic) text classification (TC henceforth) for Polish. The main objective of research was evaluation of effectivity and accuracy of different classification approaches for flexional language. Different settings of experiments included various preprocessing schemas, weighting and similarity evaluating algorithms. As Polish an example of flexional languages, the preprocessing phase is also crucial for any further NLP operations, including TC.

Our experiments were conducted to answer the following research questions:

1. Has the morphological complexity of a language an impact on the classification task effectivity and performance?
2. Will the reduction of vector dimensions (text features) to nouns and verbs contribute to improving the performance and accuracy of classification for Polish?
3. Will using a threshold or a stoplist have any substantial influence to the classification effectivity?

The main goal of the described research is to evaluate whether classification rules appropriate for English (and any other morphologically simple language) can be applied for classification of texts of more complex languages with similar effectivity and accuracy. Once positively proved it could contribute in simplifying classification process for complex languages, thus decreasing operational and time costs of such operations.

In order to answer these questions, we organized a series of experiments on Polish TC. As most of NLP techniques are originally designed for English, there is a strong need to develop new tools and solutions for languages of more complex morphology than English. During our experiments, we evaluated effectivity of classification based on Part-of-Speech features, concerning especially nouns and verbs. All pre-processing, processing and classification tasks were performed using CLARIN-PL NLP infrastructure[1] (Walkowiak, 2016). It provides well developed tools for processing Polish language[2], i.e. POS tagger - *Morpho-syntactic tagger* (Radziszewski, 2013), features extractor and

[1] Clarin-Pl: http://clarin-pl.eu, last accessed 10.08.2017

[2] There are also developed and under development NLP tools for other languages too.

Walkowiak, T. and Malak, P.
Polish Texts Topic Classification Evaluation.
DOI: 10.5220/0006601605150522
In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2*, pages 515-522
ISBN: 978-989-758-275-2

filters. Extensive introduction into CLARIN-PL services, infrastructure, work-flow and algorithms is presented in (Walkowiak, 2016).

The rest of the paper is structured as follows. Section 2 introduces the state of the art in texts classification in general and for Polish in particular. In section 3 a short description of complex Polish morphology is introduced. Section 4 outlines the Polish press news corpus that was used for classification evaluation. Section 5 presents an overview of the different experiments, while in section 6 we present results. Finally, the conclusions draw the main findings, and future work directions are introduced.

2 STATE OF THE ART

Texts classification is well recognized for English (Manning, Raghavan and Schütze, 2008), there are also considerable attempts of text classification for other languages. There are, also, attempts of TC for Polish. In (Przybyła, 2015) authors present compareson of NLP tools for English and give a list (smaller than for English language) of relevant tools for Polish. A comprehensive discussion of categorization task, and its implementation possibilities for Polish language in presented in (Zadrozny and Kacprzyk, 2006). In (Zadrożny et al., 2013) authors also developed a concept of in-depth classification. Another overview of categorization methods for Polish is presented in (Zak and Ciura, 2005). Authors compare *Naïve Bayes* (Hastie et al., 2013) and tf-idf *Roccio* classifiers (Joachims, 1997), used for Polish TC in a system for processing job advertisements. Naïve Bayes classifiers performed 2 percentage points better than tf-idf based *Roccio*. There is also (Bukowski, 2017) describing Tagnet API for Polish TC. Tagnet system uses tags organized in graph structure and offers 87% of overall accuracy. It supports, currently, five thematic categories. Also (Ciesielski et al., 2012) presents a novel method of Polish TC on the base of Wikipedia resources with application of distant supervision.

Finally, (Piskorski and Sydow, 2005) describes preliminary results of research on deploying linguistic features for classification of Polish texts. Authors discuss the impact of lemmatization and term-selection strategies based on named entities. Using canonical centroid-based classification with tf-idf vector-space model they conducted a series of classification experiments on a corpus of circa 12000 press news divided into eight categories different in topics (economy, local news, culture, science, legal affairs, sport, journalism, words news). The described experiments used only six distinct preliminary settings, i.e.: raw text terms (baseline), lemmatized terms, lemmatized with stopwords removal, lemmatized with removing of unknown words, lemmatized with removing of first names. The best results, with accuracy equal to 86.29% were obtained for the lemmatized terms and for lemmatized terms with removed first names. The baseline accuracy of this research was 85.93%. Authors also observed decreasing of classification accuracy caused by removing stopwords.

However, none of the works mentioned above answers the questions we are challenging in our research. There is no such extensive evaluation of different classifiers, weighting and matrix reduction methods. Thus, we believe, our research meets the novelty criterion and conclusions can be useful for future TC systems for Polish.

3 POLISH LANGUAGE MORPHOLOGY

During our experiment we evaluated affectivity of text classification based on bag-of-words and on Part-of-Speech approaches. The PoS approach concerned nouns, adjectives and verbs. Complex morphology of Polish language, and other flexional languages, makes NLP operations more difficult than for English. Polish language morphology is determined by extensive use of pre- and suffixes to a stem and weakly constrained word order (Eder, Piasecki & Walkowiak, 2017). Suffix usually changes the part-of-speech of the stem, making conjugation and declension systems very complex. Comprehensive descriptions and explanations of Polish grammar and morphology are given in (Feldstein, 2001; Swan, 2017; Jagodziński, 2017), while (Malak, 2013) refers to influence of Polish morphology on Information Retrieval tasks. One of substantial NLP operations step is grammatical normalization. The aim of that process is to deliver unambiguous version of a text in terms of graphical representation. Normalization concerns deriving a base form of a word from any grammatically correct form of the word. For Polish the challenge is its inflection and conjugation – a flexions of verbs and nouns. A short description of those two parts of speech is provided in following lines.

3.1 Verb

There are eleven main classes of verb conjugation in

Polish language, and some irregular verbs as well. Additionally, verbs conjugate according to person and tense while verb tenses are additionally modified by aspect. Stem of the verb is conjugation sensitive, and changes according to person, tense and aspect, like for example for irregular verb być – to be (Słownik SJP – an digital dictionary of Polish):
być, bądź, bądźcie, bądźcież, bądźmy, bądźmyż, bądźże, będą, będąc, będąca, będącą, będące, będącego, będącej, będącemu, będący, będących, będącym, będącymi, będę, będzie, będziecie, będziemy, będziesz, bycia, byciach, byciami, bycie, byciem, byciom, byciu, byli, byliby, bylibyście, bylibyśmy, byliście, byliśmy, był, była, byłaby, byłabym, byłabyś, byłam, byłaś, byłby, byłbym, byłbyś, byłem, byłeś, było, byłoby, byłobym, byłobyś, byłom, byłoś, były, byłyby, byłybyście, byłybyśmy, byłyście, byłyśmy, byto, jest, jestem, jesteś, jesteście, jesteśmy, niebędąca, niebędącą, niebędące, niebędącego, niebędącej, niebędącemu, niebędący, niebędących, niebędącym, niebędącymi, niebycia, niebyciach, niebyciami, niebycie, niebyciem, niebyciom, niebyciu, niebyć, są

or for regular verb brać – to take:
brać, bierz, bierzcie, bierzcież, bierze, bierzecie, bierzemy, bierzesz, bierzmy, bierzmyż, bierzże, biorą, biorąc, biorąca, biorącą, biorące, biorącego, biorącej, biorącemu, biorący, biorących, biorącym, biorącymi, biorę, brali, braliby, bralibyście, bralibyśmy, braliście, braliśmy, brał, brała, brałaby, brałabym, brałabyś, brałam, brałaś, brałby, brałbym, brałbyś, brałem, brałeś, brało, brałoby, brałobym, brałobyś, brałom, brałoś, brały, brałyby, brałybyście, brałybyśmy, brałyście, brałyśmy, brana, braną, brane, branego, branej, branemu, brani, brania, braniach, braniami, branie, braniem, braniom, braniu, brano, brany, branych, branym, branymi, brań, niebiorąca, niebiorącą, niebiorące, niebiorącego, niebiorącej, niebiorącemu, niebiorący, niebiorących, niebiorącym, niebiorącymi, niebrana, niebraną, niebrane, niebranego, niebranej, niebranemu, niebrani, niebrania, niebraniach, niebraniami, niebranie, niebraniem, niebraniom, niebraniu, niebrany, niebranych, niebranym, niebranymi, niebrań.

3.2 Noun and an Adjective

Also Polish declension, despite of being quite complex, offers quite a lot of irregularities. Polish nouns and adjectives are inflected by cases, genders and number. There are seven cases, two number classes, and three main genders: masculine (with subclasses: personal in sing., non-personal animate, non-personal inanimate in pl.) feminine, and neutral. Cases combined by number and gender create seventeen declension types. Noun declension inflects not only suffixes, but it may also derive the root, e.g. *człowiek* (a man), N. sg. to *ludzie*, N. pl. All grammatical forms of noun człowiek (a man) are as follow:
człowiek, człowiecze, człowieka, człowiekiem, człowiekowi, człowieku, ludzi, ludziach, ludzie, ludziom, ludźmi.

It its worth to mention that personal names in Polish are also subject to declension rules.

Due the complex morphology of Polish language using a stemmer for Polish texts classification task is not the best choice. There are few stemmers for Polish, though even for classical Information Retrieval task stemmed does not appear to increase the effectivity of pre-processing tasks. Much better, in subject (topic) or stylometric classification, performs lemmatizing approach, transforming any word in text into its basic, grammatical form – a lemma, i.e. nominative singular for nouns, for example.

4 POLISH TEXTS COLLECTION

A corpus of press news has been prepared for the research. Each of the document was manually categorized according to at least two, out of three classification systems, i.e.: internal categories of press agency, subject categories and IPTC subject codes (News Categories Taxonomy for the Media: https://iptc.org/standards/subject-codes/).

For the classifiers training we composed a sub-corpus, consisting of documents assigned to one of the following five classes, of comparable sizes in documents noumbers: diplomacy: 1082 docs., sport: 1823, disasters and accidents: 1052, economy, business and finances: 1522 and transportation: 1085. Altogether there are 6564 documents. Average size of a document was 267 words. As a main class indicator, internal categories of press agency were used for their descriptive and comprehensive contents. Over documents from this sub-corpus we conducted a series of TC experiments described in the next section.

5 EXPERIMENTS, RESULTS AND EVALUATION

The documents from training corpus were used to

compose training matrix, composed of feature vectors generated from the corresponding documents. Part of Speech tagging, accompanied by lemmatization, were performed over the training corpus texts. Lemmatization process on the base of morpho-syntactic tagger was performed and also confronted with chosen TC approaches. All the texts were send to WCRFT2, a morphosyntactic tagger for Polish, which joins Conditional Random Fields (*CRF*) and tiered tagging of plain text, for POS tagging, then *Fextor* and *Featfil* – another CLARIN-PL services - were used in order to prepare features – document matrix.

5.1 Feature-Document Matrices

Feature vectors (representation of a document) were generated using a standard 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. CLARIN-PL tagger allows to choose from 36 distinct grammatical features. The features are in accordance to The National Corpus of Polish Cheatsheet[3] (Przepiórkowski, Buczyński & Wilk, 2011). Expect of single features there is possibility of extracting their combinations as bigrams and trigram, like:

3-gram: *adv_adj_interp* – covering adverbs, adjectives and punctuation,

2-gram: *ger_adv* – covering gerund (infinitive) and adverbs.

Other research, supported partially by CLARIN-PL, proved high discriminative power of grammatical bi- and trigrams in categorization task (Maryl, 2016). Bigrams and trigram was also used in order to cluster XIX century Polish novels (Eder, Piasecki and Walkowiak, 2017), where they proved higher accuracy in stylometric analyses than other individual grammatical features.

Once generated and extracted the text features was used to build feature-document matrix on the basis of feature frequency. In respect to experiment design also two word-document matrices were prepared, too. One of them consisted of grammatical word forms and one of lemas.

The base form matrix was then used in classification run, we referred to as *baseline* run.

[3] The full list of grammatical categories distincted in NKJP is available at: http://nkjp.pl/poliqarp/help/ense2.html#x3-40002.2

5.2 Matrices Dimensions Reduction

We also conducted vectors dimensions' reduction by applying the following techniques:

- the POS filtering,
- a general stopword list,
- a threshold on the base of term frequency.

As for POS filtering, we used nouns and verbs lemmas filter. The PoS filters were used in single and mixed modes, i.e. only nouns were included in classification runs, then only verbs and, as the third setting, nouns and verbs together. Savoy (2006) proves the efficiency and accuracy of light stemming for French, Portuguese, German and Hungarian languages in context of Information Retrieval (IR). This approach was also tested for Polish (Malak, 2013). As IR and TC rely to some extend on tokens matching we decided to adopt Savoy's approach, called *light stemming*, to text classification task.

A stoplist consisted of ca. 800 most frequent words, derived from the whole 500.000 tokens corpus of press news, was used in some experiments. The stop words were the most frequent words and their grammatical forms, added in respect to complex Polish morphology.

There were also two levels of threshold applied: at a hundred and at a thousand of most frequent terms. We wanted to observe here if using threshold will affect accuracy of text classification, as its influence for efficiency is obvious.

5.3 Term Weightings

For weighting the training matrix we used classical tf-idf approach. As (Fuhr et al., 2007; Chubak and Shokouhi, 2004; Ngoc et al., 2012) state that a *Lnu.ltu* version of *tf-idf* term weighting results in better clustering accuracy than other weighting schemes, so we also tested this approach for Polish TC. By default we used *tf-idf* implementation, where the *tf* is normalized by *max*, while we also prepared implementation following Genism[4] (normalization to length 1 after *idf*), aforementioned *ltu* weighting and OKAPI weighting of terms (Manning, 2009).

5.4 Classification Runs

The study was performed according to the stratified k-fold cross-validation (with 4 folds) (Hastie et al., 2013). Word-document and feature-document vector

[4] https://radimrehurek.com/gensim/

matrices were used in classification runs. The following algorithms were used in our experiments:

- Multilayer Perceptron (MLP) (Hastie et al., 2013),
- Logistic Regression (Hastie et al., 2013),
- Decision Tree (Hastie et al., 2013),
- Random Forest (Breiman, 2001),
- Linear SVM with elastic net penalty learned by stochastic gradient descent (SVM_en_SGD) (Tsuruoka, Tsujii and Ananiadou, 2009),
- SVM (with RBF kernels) (Hastie et al., 2013).

The text classification work-flow consisted of the following processing operations:

1. testing different classifiers on grammatical and normalized forms of words,
2. applying dimensions' reduction by POS and quantitative filtering or removing stop words,
3. applying different weighting methods.

Altogether we conducted circa 216 experiments in different settings of features selected, classifiers used and reduction methods applied. Table 1. presents settings of chosen key runs.

6 RESULTS

During our TC experiments we achieved classification accuracy ratio between 74% and 95%. Evaluation was made in accordance to manual classification meta data provided by press agency to each of analysed article. Each automatic classification run results were compared to this mentioned manual class adjustments.

We achieved surprisingly good results for a combination of baseline run with linear SVM, without any weighting. This run accuracy was 89%. From all different classification runs the linear SVM appeared to be the most accurate method for Polish texts classification.

As press news are a special kind of language messages, they represent good quality of texts, which has influence into classification process. Indeed, applying other TC settings leaded us to improve text classification accuracy only for 6 percentage points – up to 95%, also for linear SVM. Despite of additional TC settings applied for individual classification runs the best results were achieved always for words base forms (lemmas).

POS filtering and nouns filtering have no discriminative influence on TC effectivity, but contributes to decreasing computational and system costs. From the other hand, using a general stoplist decreased accuracy of TC, despite of using weighting and despite of used TC approach. Threshold at 100 decreased any TC approach accuracy for about 8-10 percentage points, while a threshold at 1000 performed equally good as best TC settings.

Weighting did not appear to have any substantial influence into classification accuracy. But this result needs further confirmation by running TC for other kinds of texts than press news.

The most promising TC settings are presented in Table 2.

Table 1: Description of different features settings used in TC experiments.

Setting	Features of texts	Vertical dimension
Baseline	grammatical word forms, no reduction	109,457 features, 3.5 GB matrix
Lemmatized	base forms, no reduction	46,571 feat., 1.6 GB matrix
POS filtering	base forms, only nouns and verbs	16,157 features
lnu weighting	POS filtering + lnu weighting	16,157 features
okapi weighting	POS filtering + OKAPI weighting	16,157 features
Nouns	base forms, only nouns	14,148 features
baseline threshold 100	baseline + 100 the most frequent words limit	100 features
lemmas threshold 100	base forms + 100 the most frequent lemmas limit	100 features
baseline threshold 1000	baseline + 1000 the most frequent words	1000 features
lemmas threshold 1000	base forms + 1000 the most frequent lemmas limit	1000 features
base stoplist	base forms, stop words removal	46,281 features
baseline + stoplist	grammatical word forms, stop words removal	108,828 features

Table 2: Effectivity of chosen experiment settings.

No.	Setting	Accuracy
1.	Baseline, SVM_en_SGD	89%
2.	Baseline, SVM	89%
3.	lemmatized, tf-idf, SVM_en_SGD	93%
4.	POS filtering, if-idf genism, SVM_en_SGD	95%
5.	POS filtering, ltu, SVM_en_SGD	95%
6.	POS filtering, Okapi, SVM_en_SGD	94%
7.	POS filtering, tf-idf, threshold 100, SVM_en_SGD	87%
8.	POS filtering, tf-idf, threshold 1000, SVM_en_SGD	95%
9.	orth stoplist, ltu, SVM_en_SGD	94%
10.	nouns, ltu, SVM_en_SGD	95%
11.	orth stoplist, ltu, threshold 100, SVM_en_SGD	77%

7 CONCLUSIONS

Despite of Polish language complexity, if we deliver properly big corpus for supervised training of classifiers, we may use typical NLP TC tools, that are appropriate for English language. Our baseline run, achieved very high accuracy, equal 89%, for linear SVM classifier, while working on original texts, without any preprocessing or normalization. Similar research done by (Piskorski and Sydow, 2005) allowed to achieve 86.29% accuracy of classification, with 85.93% for a baseline run. We assume our corpus was more consistent than theirs, and probably manual classification made by press agency was more accurate. The conclusion, and an answer for our first research question is then:

The complexity of language system is of minor influence to TC process, under condition of providing big corpus of good quality texts.

Another conclusion concerns initial preprocessing operations within TC process. The grammatical normalization, in case of flexional languages, substantially decreases word- and feature-document matrices. Lemmatization increases the accuracy of TC for Polish. But then there is no substantial improvement either using a POS filtering, a threshold cut-off or a stoplist. As for very popular tool, stoplist brings poor overall effectivity improvement during TC. Filtering and threshold contribute to little improvement in TC accuracy (about one percentage point), but can really decrease system costs of TC process, because of substantial reduction of vector dimension (about ten times smaller). If we decide to filter by POS, the best choice, according to our results, are nouns. If there is no POS tagger available, one may use threshold at thousand most frequent words for reducing the vector dimensions. Then answer for our second research question is:

Reduction of vector dimensions icreases performance of text classification for Polish, but it does not contribute in effectivity of TC.

Also applying weighting does not meet our expectations, it improved TC accuracy only for one percentage point. But if we decide to weight the terms for TC process, the best choice is *ltu* or classical *tf-idf*.

From tested classifiers, the best performant was linear SVM – due fast training and high TC accuracy. Second best was MLP classifier, which performs quite well also for smaller number of features available.

A general remark, coming from described experiments, and from other TC experiments conducted recently by authors, is the vital role of the training corpus size. It should be at least 1000 documents per class, in order to train properly a classifier.

Our experiments allow us to answer hypothesis asked at the beginning of described research. Language complexity has small influence on TC effectivity. Reduction of dimensions leads to improvement of TC accuracy in two ways, first, in

small extent it increases accuracy of classification, and in bigger extend it reduces overall system costs of TC. And the last question – using stoplist especially, and threshold has no improving influence for TC effectivity. Cut-off at high threshold contributes to a decrease in required memory and processing time.

The study examined various feature generation methods, data sets, tasks (size of data), and methods of classification. Methods examining the similarity on the basis of all the words from texts turned out to be the most accurate.

8 FUTURE WORK DIRECTIONS

Based on achieved results we plan to develop a web based system that will allow to build online texts classification for any corpus of Polish texts with assigned classes. System will provide the most promising classification approaches, so other researchers will not need to evaluate TC methods.

Future research plans include extension of supervision to semi supervised methods. Moreover, we plan to investigate an open set classification problem. One may not forget the press news are a special kind of language messages. They are consistent. We plan then to repeat our evaluation on at least two other kinds of texts.

ACKNOWLEDGEMENTS

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

This research was supported by Polish National Science Center (NCN) under grant 2016/23/B/HS2/ 01323 “Methods and tools of corpus linguistics in the re-search of a bibliography of Polish book publications from 1997 to 2017.

This research was supported by Polish Ministry of Science and Higher Educa-tion (MNiSzW) under grant for restructuring the Institute of Library and Information Science at Faculty of Letters, University of Wrocław in years 2017-2017, no. 6674/E-344/R/2017.

REFERENCES

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.

Breiman, L., 2001. Random forests. In: Mach. Learn. 45(1): 5–32

Bukowski, L., 2017. Tagnet: the first Polish API for text classification, https://www.slideshare.net/LeszekBukowski/tagnet-the-first-polish-api-for-text-classification, last accessed 2017/05/10.

Chubak, P., Shokouhi, M., 2004. Evaluating Crawling Efficiency Using Different Weighting Schemes with Regional Crawler. In: Proceedings of IEEE 4 International Conference on Intelligent Systems Design and Applications (ISDA2004). Budapest, Hungary.

Ciesielski, K., Borkowski, P., Kłopotek M., A., Trojanowski, K., Wysocki, K., 2012. Wikipedia-Based Document Categorization. In: Bouvry P., Kłopotek M.A., Leprévost F., Marciniak M., Mykowiecka A., Rybiński H. (eds) Security and Intelligent Information Systems. Lecture Notes in Computer Science, vol 7053. Springer, Berlin, Heidelberg.

Eder, M., Piasecki, M., Walkowiak, T.: Open stylometric system based on multilevel text analysis. In: Cognitive Studies 17| Études cognitives, 2017(17). https://doi.org/10.11649/cs.1430

Feldstein, R., F.. 2001. A Concise Polish Grammar, SEELRC. Duke University.

Fuhr, N., Lalmas, M, Trotman, A., 2007. Comparative Evaluation of XML Information Retrieval Systems, Springer Science & Business Media.

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.

Jagodzinski, G., 2017. A Grammar of the Polish Language, http://grzegorj.w.interia.pl/gram/en/gram00.html, last accessed 2017/05/10.

Joachims, T., (1997) A probabilistic analysis of the Roccio algorithm with tfdf for text categorization. In: Fisher, D.H. (ed.) ICML, 143–151.

Malak P.: The Polish Task within Cultural Heritage in CLEF (CHiC) 2013. Torun Runs, In: Forner P., Navigli R., Tufis D. (red): CLEF 2013 Evaluation Labs and Workshop Working Notes, 23 - 26 September, Valencia – Spain.

Manning, Ch. D., Raghavan P., Schütze H., 2009. Introduction to Information Retrieval, Cambridge University Press.

Maryl, M. (2016). *Tworzenie typologii gatunków piśmiennictwa multimedialnego na przykładzie blogów – propozycja metodologiczna* w: Metody badań online. Piotr Siuda. Gdańsk: Wydawnictwo Naukowe Katedra, ss. 360-398

Ngoc et all., 2012. Advanced Methods for Computational Collective Intelligence, Springer.

Piskorski, J., Sydow, M., 2005. Experiments on classification of Polish Newspaper. In: Archives of Control Sciences, Special issue on Human Language Technologies as a challenge for Computer Science and Linguistics Part II, editor: Z. Vetulani, 15: 613-625.

Przepiórkowski A., Buczyński A., Wilk J., *The National Corpus of Polish Cheatsheet*, from http://nkjp.pl/poliqarp/help/en.html last accessed 2017/10/14.

Przybyła, P., 2012. Issues of Polish Question Answering. In: Proceedings of the first conference "Information Technologies: Research and their Interdisciplinary Applications (ITRIA 2012), Warsaw, Poland.

Radziszewski, A., 2013. A tiered CRF tagger for Polish, Intelligent Tools for Building a Scientific Information Platform. Studies in Computational Intelligence, 467: 215–230.

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

Słownik SJP.PL, from https://sjp.pl/ last accessed 2017/10/14.

Swan, O. 2017. Polish Grammar in a Nutshell, http://polish.slavic.pitt.edu/firstyear/nutshell.pdf, last accessed 2017/08/10.

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.

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.

Zadrozny S., Kacprzyk J., 2006. Computing with words for text processing: An approach to the text categorization, In: Information Sciences, 176: 415–437.

Zadrożny, S. , Kacprzyk, J., Gajewski, M., Wysocki, M., 2013. A novel text classification problem and its solution, In: Czasopismo Techniczne. Automatyka, R. 110, z. 4-AC 2013, 7-16.

Zak I., Ciura M., 2005. Automatic Text Categorization, In: Information Systems Architecture and Technology, Szklarska Poręba, Poland.