Towards Automatic Detection of Correct Domain Words in OCR Texts from Polish Digital Libraries

Filip Graliński*, Rafał Jaworski*, Piotr Wierzchoń†

* Department of Natural Language Processing
Adam Mickiewicz University in Poznań, Poland
{filipg, rjawor}@amu.edu.pl

†Institute of Linguistics
Adam Mickiewicz University in Poznań, Poland
wierzch@amu.edu.pl

Abstract

This paper presents an experiment on automatic classification of words, conducted on textual data coming from the Polish Digital Libraries. The main goal was to implement an algorithm which would aid manual extraction of domain-specific vocabulary from raw texts. In this scenario, the electronic format of input texts was a result of optical character recognition, which did not present satisfactory quality. Furthermore, as the main goal of the vocabulary extraction task was the identification of words not listed in any of the well-known Polish domain dictionaries, all dictionary words were filtered out from the input texts. Thus, the only words remaining were either OCR errors or valuable, rare specialist words not present in dictionaries. As expected, the valuable words were overwhelmed in terms of quantity by the erroneous tokens. The goal of the classification algorithm developed in this research is to predict whether a given word is a correct word in a given domain.

1. Introduction

Recent advances in the area of digital humanities see numerous projects dealing with historical texts. In this study, the authors are faced with the task of compiling a set of Polish domain dictionaries, containing words confirmed in various publications from the interwar period. As traditional lexicographical work may be tedious and excessively time-consuming, the authors sought solutions from the area of Natural Language Processing and Machine Learning, in order to improve the efficiency of the process.

Section 2. of the paper presents related work in the area of digital lexicography. Section 3. describes the input data for the lexicographical process. The experiments on automatic classification of correct words were organised as a challenge, which is described in Section 4.. The best experimental results are presented in Section 5., while Section 6. lists the conclusions.

2. Related work

This section gives a brief insight into similar research in the area of digital lexicography. It presents algorithms for terminology extraction, as well as automatic thematic classification of words and documents.

2.1. Terminology extraction

The input of the process of terminology extraction is a monolingual or bilingual corpus of significant size. The output is a list of candidate words and short phrases, which are to be checked manually by annotators. From the technical point of view, the extraction can be based either on linguistic rules or on statistics. The rule-based approach requires the preparation of formal grammars, which are used directly for the extraction of those phrases in a corpus which constitute a syntactically correct phrase, e.g. a noun phrase, a verb phrase or a prepositional phrase. On the other hand, in the statistical approach, the most frequent words and collocations found in a large corpus are retrieved. It is not guaranteed that the resulting sequences of words will constitute proper phrases in a linguistic sense, as the extraction process relies solely on counting word and phrase frequencies. It is therefore not uncommon to extract spurious phrases such as “make the”.

Nevertheless, in a standard scenario, terminology extraction allows the identification of a considerable quantity of valuable phrases. Hence, it is commonly used as an auxiliary tool for preparation of linguistic resources.

The papers (Seljan and Gašpar, 2009) and (Seljan et al., 2013) give an insight into the process of terminology extraction and present an experiment evaluating various extraction tools applied to Croatian and English texts. The tools included:

- NooJ – rule-based, relying on a set of regular grammars (Donabédian et al., 2013);
- Multi Term Extract – statistical, a part of the SDL Trados environment;
- PhraseFinder – a hybrid, also developed by SDL.

The results of the experiment showed that statistical tools did not outperform rule-based algorithms significantly. According to the authors, this effect could have been caused by the insufficient size of the input corpus, as statistical methods are expected to provide better results for large input data. However, these results also proved that the usage of linguistic knowledge can lead to satisfactory results in the process of terminology extraction.

Another article: (Seppälä et al., 2012) presents a rule-based approach to extracting paradigmatic properties of French adjectives from a large lexicographic dictionary.

Although this is not classic terminology extraction (as the input data is already a dictionary, not a corpus), the described process constitutes an interesting example of lexicon compilation. The core idea is to apply automatic rules to input data and have manual annotators check the output. The annotators analyse the most common errors and suggest improvements to the rules. When the suggestions are implemented, the entire process starts over. This method was found to lead to satisfactory results while minimising the human effort.

2.2. Thematic classification

The paper (Lavelli et al., 2002) describes the idea of using thematic classification as a lexicon building technique. The building process is iterative and relies on bootstrapping. The starting lexicon L_0 solely consists of terms manually assigned to several predefined categories. This first lexicon is used as a training corpus for an automatic thematic classifier, which is then applied on a new, unseen corpus. Terms accepted by this classifier are then added to the terms of L_0, resulting in L_1 ($L_0 \subseteq L_1$). For thematic classification in each bootstrapping step, the authors use custom term indexing algorithms.

The results obtained using this technique demonstrate the high precision of the extracted terms. The authors report up to 76% of the terms verified as correctly classified in a specific domain by human judges. However, the reported recall of 5–20% indicates that many more terms could have been extracted from the input corpora. This might be because the method focuses around the starting set of terms. In our work, on the other hand, it is crucial to extract specialist, sparse terms which might appear only once in the input corpora.

Some experiments with the thematic classification of Polish documents dating from the interwar period have already been conducted (Borchmann et al., 2015). Although the techniques developed in this experiment have been designed to work at the document level, they introduce a novel algorithm for thematic classification and identification of document categories. The input data for the algorithm was a set of Polish documents with manually defined (tentative) category in the Universal Decimal Classification (UDC) format. This classification, introduced in the early 20th century, serves the purpose of ordering library collections, and has nine main classes. The key idea was to perform content-based clustering on these documents by applying the following steps:

1. vectorising the documents with a technique using the idea of TF-IDF;

2. applying the spectral clustering method.

After this operation, a set of documents was obtained and assigned to categories and to clusters. Analysis of this data proved that documents assigned to some of the manually defined categories (e.g. Biology and Botany) do not differ significantly in content. Some categories suggested by UDC were merged to obtain a list of actual categories that are represented in the input texts. In this paper, we decided to use these categories.

3. Input data

The corpus of input data for the automatic classification was taken from documents catalogued in the Polish Digital Libraries. All the documents dated from the interwar period (1918–1939). As the main goal of the project is to compile domain-specific dictionaries, a list of domains of interest was necessary. To prepare this, we performed manual selection of documents, whose domain was not in doubt. For instance, we assumed that a collection of medical scientific journals, such as *"Warszawskie Czasopsimo Lekarskie" (The Warsaw Medical Journal*), would belong to the domain of medicine. Using these preliminary manual assignments, we managed to find large text collections belonging to specific domains (see Table 1).

Each of the sources was then processed in order to extract candidate lemmas. The processing pipeline is presented in Figure 1. Firstly, the source texts of a given domain are tokenised by whitespace. The resulting list of tokens is sorted and deduplicated, in order to prepare the list of raw words. Next, the raw words undergo a normalisation process which essentially consists of two stages: lower-casing and orthography correction. The second procedure is executed to correct some of the most common OCR errors (e.g. *ą* misrecognised as *q* between consonants). Additionally, a so-called diachronic normalisation is performed, which aims to alter words from interwar texts to conform to modern spelling rules (introduced in the 1936 spelling reform). Though the differences in Polish spelling rules between the interwar period and modern times might seem cosmetic, they affect a large percentage of all words. An example of such a difference can be seen in the word ending *-ria*, which in the interwar period was often spelled *-rja*. Diachronic normalisation helps to identify the same words regardless of the spelling rules they follow.

Lower-cased and normalised words undergo lemmatisation, using the LemmaGen software (Jursic et al., 2010). LemmaGen is trained on a large Polish morphological dictionary and, most importantly, is capable of lemmatizing unknown words by analogy. For example, given a non-existing Polish word "xxxowego", LemmaGen generates its lemma as "xxxowy". In this example, the ending "-owego" is recognized by the lemmatizer as a typical ending for adjectives in the genitive case, which in nominative case have the ending "-owy", e.g. "kolorowego" (of colorful) → "kolorowy" (colorful). This feature of the lemmatizer is crucial when working with OCR tokens, which typically contain various spelling errors and are therefore not found in any dictionary, let alone in a morphological one.

The last step in the pipeline is the filtering out of lemmas which are contained in at least one of the well-known Polish electronic dictionaries, including PoliMorf (Woliński et al., 2012). Thus, the only lemmas remaining are either erroneous, or are valuable, specialist domain words not included in common dictionaries.

The task of identifying the correct words among numerous erroneous tokens is given to a group of human annotators. Their work is aided by a custom-made web-based application. The annotators are given 500 tokens in

Source	Lines	Words	Characters
astronomy	291 740	1 431 552	9 636 232
linguistics	137 483	1 385 296	9 841 025
mathematics	177 069	980 882	6 578 932
medicine	4 129 461	38 431 502	269 797 823
military	1 018 290	7 531 491	53 317 472
occultism	537 558	5 085 710	35 129 981
religion	6 381 426	48 352 193	344 208 129
sport	1 476 641	9 986 787	69 276 368

Table 1: The sizes of source texts.

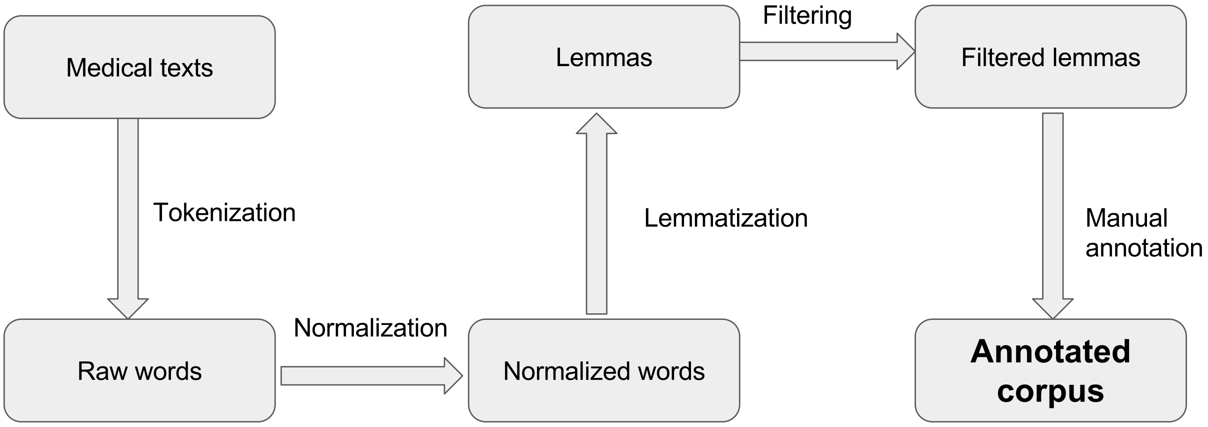

Figure 1: Source processing pipeline.

a single batch that contains tokens extracted by the above pipeline from source texts in a specified domain. They are then asked to decide which tokens are correct words in that domain – that is, which words can be entered in a specialist dictionary for the domain. The annotators are instructed not to use any external source of information to verify the correctness of words. Their analysis at this point is assumed to be superficial and based solely on human intuition, the ability to use simple heuristics and some general knowledge. Annotators are instructed to bias their decisions in doubtful situations towards the acceptance of suspicious tokens. The final annotation will be performed by experts in each of the domains.

All accepted lemmas carry information about the normalised words from which they were derived, and each normalised form carries information about its raw token versions. Table 2 presents the lemma *abiotroficzny* (Latin: *abiotropicus*), accepted as a correct word in the medical domain. Numbers in parentheses represent the frequencies of the forms in source texts.

4. The sane words challenge

A series of experiments was performed to determine whether the first round of superficial human annotation can

Lemma	Normalised	Raw words
abiotroficzny (9)	abiotroficzne (3)	abiotroficzne (2) Abiotroficzne (1)
	abiotroficzny (4)	abiotroficzny (2) Abiotroficzny (2)
	abiotroficznych (2)	abiotroficznych (2)

Table 2: Example of an accepted lemma.

be automated. This task is viewed as a binary classification problem, where the input data consists of:

- a literal token t;
- the reported frequency f of the token in source texts;
- the name of the domain d.

The output should be a simple yes/no answer, interpreted as "the token t can/cannot be considered as a specialistic word in the domain d".

The experiments were performed using the *Gonito.net* platform (*Gonito.net* is a platform for machine learning competitions (Graliński et al., 2016)). Thus, the task of preparing an automatic classifier was presented as a challenge to a group of researchers (see `http://gonito.net/challenge/sane-words`), who competed (and cooperated) to achieve the best results and could easily exchange their findings.

For the sake of the challenge, appropriate training, development and test sets were prepared. The data came from the results of the preliminary human annotation, performed on 66,431 tokens (mainly from the domain of medicine) annotated as correct or incorrect. 44,344 tokens chosen at random from the above set constituted the training set. The percentage of accepted tokens in the training set was only 4.0%. The development and test sets contained 11,026 and 11,061 tokens, with acceptance rates of 3.7% and 3.6% respectively.

The metric for scoring classifiers submitted as solutions to the "sane words challenge" was the F_2 score (F-score biased towards recall).

5. Best results

Two solutions submitted by researchers from our team will be presented here. The first was based on machine learning methods offered by the Vowpal Wabbit software (Langford et al., 2009), while the second relied on a simple neural network.

5.1. Vowpal Wabbit

The solution to the "sane words challenge" based on Vowpal Wabbit is currently the best-performing classifier, reaching an F_2 score of **0.41** on the test set. Its implementation is based on a logistic regression algorithm with the following features:

- character bigrams;
- token prefixes and suffixes of length up to 4;
- whether the word is accepted by the Aspell spellchecker (binary feature);
- the length of the word (in characters);
- number of vowels in the word;
- number of occurrences of the letters q, v and x (which are absent from the Polish alphabet and occur only in a limited number of loanwords);
- number of Google results for the word as search query;
- word frequency (from training data);
- word domain (from training data).

Despite testing various other features, such as spellchecking with a German spellchecker or introducing higher n-grams of letters, the results did not improve.

The solution and its output is available at Gonito.net {879a88} (in case you are reading a physical copy of this paper, go to `http://gonito.net/q` and enter the reference number there).

5.2. Neural network

A very simple neural network was trained with the Keras deep learning library[1] (Gonito.net reference {285709}). The main component of the network was a character-level LSTM working on 4-dimensional character embeddings. The only other feature used was the word frequency. Even though much less feature engineering was used than in the approach based on logistic regression, the result was comparable (an F_2 score of **0.40**). Making the neural network more complex (more layers, use of dropout) did not improve the result (which is not surprising, as the amount of data is rather small for deep learning).

Interestingly, the result based on the neural network and the Vowpal Wabbit solution achieved a very similar score, despite the differences between these approaches. This might indicate that the F-score of 0.4 in this scenario is a boundary which can not be improved significantly, let alone easily.

[1] https://keras.io

Acknowledgements

Work supported by the **Polish Ministry of Science and Higher Education** under the **National Programme for Development of the Humanities**, grant 0286/NPRH4/H1a/83/2015: “50 000 słów. Indeks tematyczno-chronologizacyjny 1918-1939”.

6. Conclusions

This paper presented a series of experiments on automatic classification of correct, domain-specific words in a set containing a considerable number of erroneous OCR tokens. The results of the experiments enabled the preparation of a binary classifier capable of achieving an F2-score of 0.41. A classifier achieving this score is sufficient to be used in the scenario, where a large portion of tokens is first filtered automatically and then presented to human annotators, who perform the decisive classification. This will enable a significant boost in the efficiency of the terminology extraction process, as humans would only annotate tokens exhibiting a high probability of acceptance.

As the manual annotation of tokens is an ongoing process, future work plans include repeating the “sane words challenge” with new, augmented training data.

7. References

Borchmann, Łukasz, Filip Graliński, Rafał Jaworski, and Piotr Wierzchoń, 2015. A semi-automatic method for thematic classification of documents in a large text corpus. In *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH)*. Institute of Computer Science of the Polish Academy of Sciences.

Donabédian, A., V. Khurshudian, and M. Silberztein, 2013. *Formalising Natural Languages with NooJ*. Cambridge Scholars Publishing.

Graliński, Filip, Rafał Jaworski, Łukasz Borchmann, and Piotr Wierzchoń, 2016. Gonito.net – open platform for research competition, cooperation and reproducibility. In António Branco, Nicoletta Calzolari, and Khalid Choukri (eds.), *Proceedings of the 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*. pages 13–20.

Jursic, Matjaz, Igor Mozetic, Tomaz Erjavec, and Nada Lavrac, 2010. LemmaGen: Multilingual Lemmatisation with Induced Ripple-Down Rules. *J. UCS*, 16(9):1190–1214.

Langford, John, Lihong Li, and Tong Zhang, 2009. Sparse online learning via truncated gradient. In *Advances in neural information processing systems*.

Lavelli, Alberto, Bernardo Magnini, and Fabrizio Sebastiani, 2002. Building thematic lexical resources by term categorization. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’02. New York, NY, USA: ACM.

Seljan, Sanja, Ivan Dunđer, and Angelina Gašpar, 2013. From digitisation process to terminological digital resources. In *Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention on*. IEEE.

Seljan, Sanja and Angelina Gašpar, 2009. First steps in term and collocation extraction from english-croatian corpus.

Seppälä, Selja, Alexis Nasr, and Lucie Barque, 2012. Extracting a semantic lexicon of french adjectives from a large lexicographic dictionary. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval ’12. Stroudsburg, PA, USA: Association for Computational Linguistics.

Woliński, Marcin, Marcin Miłkowski, Maciej Ogrodniczuk, Adam Przepiórkowski, and Łukasz Szałkiewicz, 2012. PoliMorf: a (not so) new open morphological dictionary for Polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*. Istanbul, Turkey: ELRA.