RetroC – A Corpus for Evaluating Temporal Classifiers

Filip Graliński*, Piotr Wierzchoń†

Adam Mickiewicz University
*Faculty of Mathematics and Computer Science / †Institute of Linguistics
*Umultowska 87, 61-614 Poznań, Poland / †al. Niepodległości 4, 61-874 Poznań, Poland
filipg@amu.edu.pl, wierzch@amu.edu.pl

Abstract

We present a corpus for training and evaluating systems for the dating of Polish texts. A number of baselines (using year references, knowledge of spelling reforms and birth years) are given for the temporal classification task. We also show that the problem can be viewed as a regression problem and a standard supervised learning tool (Vowpal Wabbit) can be applied. So far, the best result has been achieved with supervised learning with 5-grams, year references and rules based on spelling reforms.

1. Introduction

In recent years, more and more historical material (such as old newspapers, books no longer under copyright and archival documents) has been digitised and made available online. Unfortunately, metadata, in particular creation/publication dates, is not always present. Moreover, old textual material is often made available on the Internet in an unstructured manner and mixed with contemporary Web texts.

The task of *automatic document dating* or *temporal text classification* consists in assigning a creation or publication date to a given text relying solely on its content – that is, without the need to use explicit metadata (Dalli and Wilks, 2006). It can be viewed as a text classification problem (which period does it come from? – with, for instance, a yearly or decadal resolution) or as a regression problem (guess the time stamp as precisely as possible treating it as a continuous value). The task can be approached using either knowledge-based methods (through knowledge of the history of the orthography of a given language or using Wikipedia or other external resources) or learning-based methods (supervised learning from a corpus of time-stamped texts).

In this paper, we present (1) the first version of *RetroC* – a publicly available corpus for evaluating and training systems for the automatic dating of Polish texts and (2) some baseline results obtained using the corpus.

In section 2. we discuss previous work and the state of the art as regards temporal classification. Section 3. presents the rationale behind the RetroC corpus, its source materials and scope. Some basic baselines and statistics are given in section 4. More advanced knowledge-based methods are discussed in section 5. and in section 6. we present some preliminary results obtained using supervised methods.

2. Previous work

Although the problem of temporal classification is of significant importance for text processing and information retrieval, as well as in terms of its numerous applications (language chronologisation, support for the digitisation of cultural heritage), relevant literature is not abundant. This is probably a result of the lack of large, freely available, resources which might be used to train and test automatic dating systems.

The research problem was first raised by Jong et al. (2005). Those authors presented an ambitious programme of using temporal unigram language models not only for the automatic dating of historic texts, but also for linking contemporary keywords with their historic variations. Since, unfortunately, there was no extensive diachronic corpus available, de Jong et al. carried out the experiment based on a fairly large but time-limited (1999–2005) corpus of Dutch-language press materials. Kanhabua and Nørvåg (2009) developed further the method of Jong et al. (2005) by applying semantic-based pre-processing (tagging parts of speech, excerpting collocations and filtering out words) and using statistical extensions of language models (word frequency interpolation, temporal entropy, the use of Google Zeitgeist). In order to learn and test the methods, a corpus of archival websites from an approximately 8-year period was used.

Evaluation of automatic dating methods was one of the objectives of the DEFT2010 workshop. To this end, a time-extensive (1800–1944), though relatively small (about 6300 texts) corpus of French newspaper texts was used. The best system created by Albert et al. (2010) obtained an F-measure of 0.338. Use was made of information about spelling reforms, birth dates of famous people and a module which learnt to chronologise vocabulary with conditional random fields.

The evaluation task was repeated during the DEFT2011 campaign. An advanced system based on information gained from external resources (birth dates, archaisms, neologisms, dates of spelling reforms) and on classification methods making use of a training corpus (classification based on the cosine distance, with modelling using support vector machines) was then constructed by Garcia-Fernandez et al. (2011).

In order to improve the automatic dating results, Chambers (2012) made use of a discriminant classifier, taking into account explicit temporal references in the dated text and parameters such as verb tense. Kumar et al. (2012) applied language models learned from Wikipedia biographies

to classify stories obtained in the Gutenberg Project, and Ciobanu et al. (2013) trained a classifier based on a Romanian corpus, containing data from five centuries, to date contemporary historical novels. (This is a more difficult task than dating, for instance, press articles, which usually refer to events that are not distant in time from their publication dates.)

Recently, Guo et al. (2015) applied various machine learning methods (e.g. SVMs) to a large dataset extracted from the HathiTrust digital library.

3. The RetroC corpus

RetroC is a Polish-language diachronic corpus, spanning two centuries (1814–2013) and intended for training and testing automatic dating systems. It is mostly based on publications available in Polish digital libraries (Wierzchoń, 2009; Graliński, 2013), plus some old textual material from other online sources.

The corpus was designed with the following goals in mind:

- to be a collection of Polish texts;
- to be large enough to enable the use of statistical methods;
- to be time-extensive – not just modern Web-based texts, but also old printed materials;
- to cover relatively short fragments rather than whole books, for which the dating task is much easier.

The first release of RetroC consists of three sets:

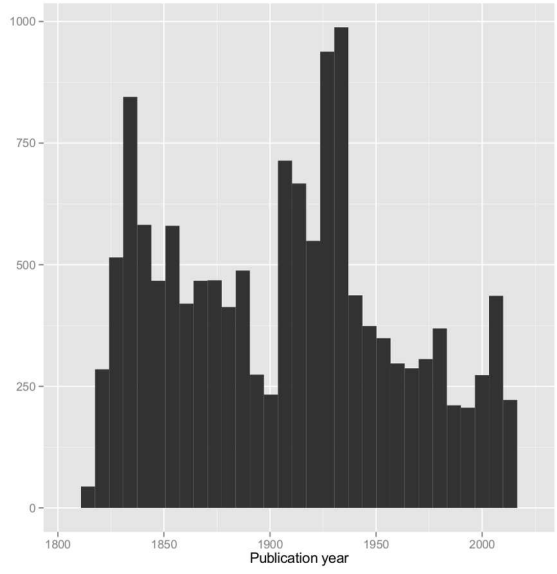
- training set (40,000 fragments),
- development set (9,910 fragments),
- test set (10,000 fragments).

Each set is composed of 500-word fragments taken from random publications (500-word portions were also used in the DEFT corpus (Garcia-Fernandez et al., 2011)). For instance, the following is a dev-set item taken from an 1855 publication from the e-library of Warsaw University[1] (which is the largest source of texts for the training and development sets):

> przeprawę. Szron ten zwiększa się w skutku przymrozków i śniegu, na czem w tych dniach zupełnie nam niebrak. Zapowiedziany Toro IV i ostatni dzieła p.n. Opisanie lasów Królestwa Polskiego i Gubernji Zachodnich CESARSTWA Rossyjskiego, już wyszedł z droku i znajduje [omitted for brevity] ubioru damskiego zastosowane, wyszły na r. 1855 nakładem i w litografji K. Romanowicza, przy ulicy Długiej Nr 578, przechodni dom na Bielańską. Nabyć ich także możua w składzie ryciu przy ulicy Senatorskiej, wdomuW 7 Neubauera. Nakładem Xięgarni Jana Breslauera, wyszła z druku powieść historyczna: Zamek Warszawski czjfli Rodzina Konrada, w 3ch tomach, przez J. N. Cżarnomskiego. Powieść ta opisuje w sposób nader zajmujący, ostatnie chwile Xięztwa Mazowieckiego i jego wcielenie do Korony. Cena exem: rs.2 k. 70 Rzadko takiego kursu sanek jak w dniu onegdajszyro, bo też dzień >był potemu, gdyż i dość mroźny, zatem pogodny i śnieg

[1] http://ebuw.uw.edu.pl

Figure 1: Number of items in the training set



As can be seen, a text in the RetroC corpus is given as it was found in the text layer of a DjVu/PDF file (with possible OCR noise and errors) – only minimal post-processing was applied (joining words separated with hyphens and new lines, removing end-of-lines and other non-printing characters, UTF-8 sanitisation). In contrast to the DEFT dataset (Garcia-Fernandez et al., 2011), dates were not removed from texts (see *1855* in the example above); this was motivated by the fact that year references are obviously a useful (though not perfect) feature for temporal classification (and we aim to use classifiers trained with RetroC to find old textual material in large Web corpora where dates are available), although it is not as important as in (Guo et al., 2015), where whole volumes, including copyright and title pages, are taken into account.

The development and test sets are balanced with respect to publication year: 50 publications per year. We were not able to find very many dev-set items for some years during World War II, hence the size of the development set is smaller than 200 (years) $\times$ 50 (texts). The development set and the test set are also balanced (as much as possible) with respect to their sources, in order to avoid the data set being overwhelmed by one large digital library. The training set is not balanced; the distribution of publication years therein is presented in Figure 1.

The development and training sets are composed of texts from the same set of digital libraries. In order to make the challenge more difficult, the texts in the test set were

Figure 2: Year references

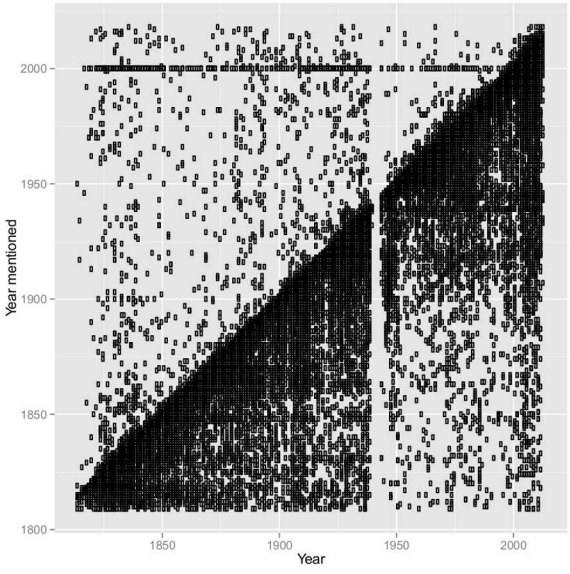
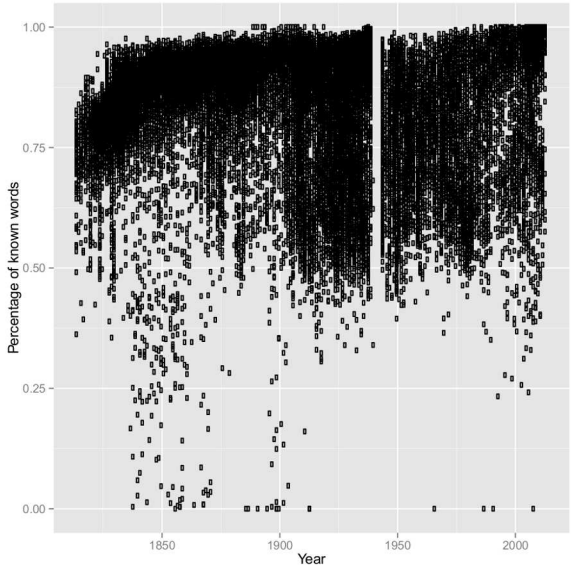


Figure 3: Percentage of known words in the training set



taken from a separate set of sources (i.e. different digital libraries).

The publication dates were taken from the metadata from the digital libraries; no manual verification was performed, and there is no guarantee that all of the dates given in RetroC are correct (we also ignore whether it is in fact a publication date or creation date that is given).

The corpus is freely available from a git repository (`git://gonito.net/retroc.git`; the target values for the test set are available in the `dont-peek` branch). The machine learning task defined along with the corpus is configured to use root-mean-square error (RMSE) as the evaluation metric.

4. A look into the corpus

4.1. Year references

A simple tool was created to extract year references from RetroC texts – it was based on a regular expression for the time span in question ± 5 (1809–2018) plus a blacklist of words which cannot follow a year reference (e.g. monetary units). Of the items in the training set 52.1% do not contain any year reference, 17.9% contain exactly one year reference, and 30.0% contain more than one. The distribution of year references vs. publication years is shown in Figure 2.

Based on observation of Figure 2, the following simple baseline for automatic dating may be proposed:

- if no year reference is found, return 1913 (median year);
- if only references to year 2000 are found, return 2000;
- otherwise, return the latest year reference other than 2000.

This method yields RMSE = 45.9 years (vs RMSE = 57.9 for the null model that always returns 1913).

4.2. OCR noise

It might be assumed that the older the text, the lower the quality of the OCR output (and the higher the probability of finding words written with obsolete orthographic conventions, archaic words, etc.). In order to check this assumption, we counted the percentage of words recognised by a modern spell-checker; see Figure 3. Unfortunately, the linear regression model with respect to the logit of these percentages yields the same RMSE as the null model (57.9).

5. Knowledge-based methods

5.1. Changes in Polish orthography

Because of the country's turbulent history, the study of Polish orthography covers a whole range of issues related to frequent reforms, geographical division, fragmentation of the academic community, the preferences of publishers and users of the language, and so on. Describing the changes in and practical application of the rules of Polish orthography is an extremely complex task. Here we will attempt to summarise this topic in brief.

At the start of the Modern Polish era, matters of orthography were taken up by Onufry Kopczyński. He reintroduced the letters *á* and *é*, although their usage was very difficult. In 1816, Alojzy Feliński rejected the use of the character *á*, while retaining *é* and *ó*. In place of the formerly used *i* and *y* he introduced *j* (*jajko* instead of *iayko*).

In 1890 the Academy of Learning set up an Orthographic Commission, which in 1891 published a set of rules prescribing:

- the forms *módz*, *biedz* (rather than *móc*, *biec*);
- *ge* rather than *gie*, for example *geografia* rather than *gieografia*;

- *Francya*, *Anglya* (not *Francja*, *Anglia*), *Maryja*, *Julia* (not *Marja*, *Julja*), *kolacya*, *dyalekt*, *dyagnoza*;
- the endings *-im*/*-ym* as alternatives to *-em*, and *-imi*/*-ymi* as alternatives to *-emi*.

These reforms met with serious objections. Consequently, in 1906, the Linguistic Commission of the Academy of Learning approved the following principles:

- *-ja* in place of *-ia*, *-ya*, but only in final syllables (e.g. *biologja*, *diatermja*);
- *gie* in place of *ge*, as in *gieografja*, *gienerał*;
- the endings *-ym*, *-im* and *-ymi*, *-imi* (alternatives to *-emi*);
- the infinitive forms *biec*, *móc*.

Further innovations followed in 1918. The changes announced in the Main Principles of Spelling included the following:

- in non-initial syllables, in foreign words, the letter *j* to be used after a consonant;
- the endings *-em*, *-emi*, *-ym*, *-ymi* to remain distinct according to the ending of the nominative;
- foreign words to be written with *ke*, *ge*, but *kie*, *gie* to be used in words felt to be native;
- *-c* to be used in infinitives of the type *biec*, *móc*.

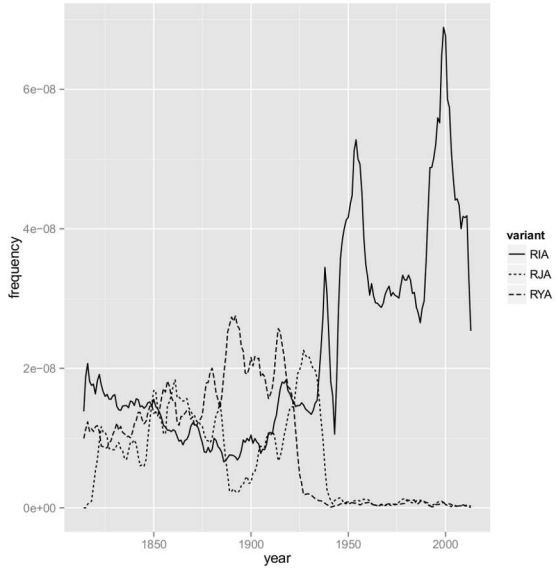
These rules were not universally adopted, and moreover they sparked controversy. Hence, in 1932 a new set of rules appeared, although these related mainly to questions of spacing. For example, it was decided that expressions consisting of a preposition and a noun would be written as a single word if they had an adverbial meaning, as *pomału*, *zamłodu*, *zbliska*, *nakoniec*.

The greatest reform of the orthographical system was enacted in 1936:

- words like *Maria* to be written with *i*, except after *c*, *s*, *z* (e.g. *Francja*, *pasja*, *diecezja*);
- inflectional endings of adjectives to be standardised as *-ym*, *-ymi*;
- *ke* in foreign words to be written *kie*, but foreign *ge* not as *gie*;
- the negating particle *nie* should not be written as one word together with participles having a verbal meaning;
- changes were made to the spellings of certain specific words, such as *brózda* → *bruzda*, *chróst* → *chrust*.

Taking into account this knowledge and graphs such as that shown in Figure 4, we decided to partition the whole time span into the three periods 1814–1918, 1918–1936 and 1936–2013, score the periods according to Table 1, and take the median year in the winning period. This method results in recall 58.8%, precision 83.1% (for recognition of one of the three periods), and RMSE = 46.6.

Figure 4: Frequency of orthographic variants



	1814–1918	1918–1936	1936–2013
-dz ending	1.0	0.5	0.0
cya	1.0	0.5	0.0
rya	1.0	0.5	0.0
rja	0.5	1.0	0.0
-emi ending	1.0	1.0	0.0
-ymi ending	0.0	0.0	1.0

Table 1: Scores for specific variants

5.2. Using external sources

External sources of knowledge, such as Wikipedia, can be used in automatic dating systems. We extracted the birth years of 210,297 people from Polish Wikipedia (cf. (Garcia-Fernandez et al., 2011)). Unfortunately, it turned out that only 12.7% of the training set contained references to any of these people, and the level of noise was quite high – 19.3% of matches appeared in fragment dated *earlier* than the corresponding birth year (even though both the first name and the surname were required for a match). It is not surprising, then, that the predictor based only on birth years is no better than the null model (RMSE = 57.4).

6. Experiments with supervised learning

Temporal classifiers can be trained with supervised learning methods (supply a collection of texts annotated with dates, and apply a machine learning classification or regression algorithm). So far, classification approaches have been used (create a separate model for each year, and for a given text assign the year for which the classifier yielded the highest probability or score); see e.g. (Garcia-Fernandez et al., 2011). Here, we will consider the regression approach: one model with the output (year) taking continuous values (given with any precision, not necessar-

	positive	negative
1.	stori	tém
2.	czym	aig
3.	wtedy	»
4.	*dash*	il
5.	”.	i5
6.	tzw	)"
7.	‘	téj
8.	ktoś	tal
9.	2009	1837
10.	1985	storj

Table 2: The features with the highest scores

method	dev	test
null model	57.9	57.7
year references	45.9	46.4
orthography	46.6	50.9
5-grams	22.0	33.5
5-grams + year ref. + ortho.	21.7	33.1

Table 3: Summary of results

ily rounded, i.e. we accept values such as 1951.440777).

We used the Vowpal Wabbit open-source learning system (Langford et al., 2009). As features we simply used lower-cased tokens and/or character pentagrams (as suggested in (Garcia-Fernandez et al., 2011)); the predicted value is $y - 1913.0$ for year y, and the number of training passes was 40. In order to account for the difference in distribution between the training set and the development/test set, we applied inverse weighting of training examples (with the log function).

Using pentagrams yielded a better result (RMSE = 22.0) than using tokens (RMSE = 22.8). This is not surprising, as character-level n-grams are more robust to OCR noise, and they are a simple yet effective substitute for lemmatisation in the case of inflected languages such as Polish. (The n-grams with the highest weights are presented in Table 2.) Using both tokens and n-grams did not improve the quality of the system (RMSE = 21.9).

A natural idea is to add the parameters discussed in sections 4. and 5. as features. Interestingly, neither adding the year extracted as described in section 4. (RMSE=21.9) nor knowledge about changes in orthography (section 5.1. RMSE=21.8) improves the quality of the system significantly. If both of these types of features are added, we obtain RMSE=21.7. A summary of the results obtained on both the dev set and the test set is presented in Table 3.

7. Conclusions and further work

We have presented RetroC, a Polish corpus for evaluating temporal classifiers, and reported initial results for certain methods. It has been shown that automatic dating can be treated as a regression problem, and that a standard machine learning tool (Vowpal Wabbit) can be used to obtain fairly good results.

For future work, we plan to implement all of the advanced classification methods known in the literature for other languages, and compare and combine them with the regression methods. Also, RetroC will be extended and a new test set will be prepared.

8. References

Albert, Pierre, Flora Badin, Maxime Delorme, Nadege Devos, Sophie Papazoglou, and Jean Simard. 2010. Décennie d'un article de journal par analyse statistique et lexicale. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*.

Chambers, Nathanael. 2012. Labeling documents with timestamps: Learning from their time expressions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics.

Ciobanu, Alina Maria, Liviu P Dinu, Octavia-Maria Sulea, Anca Dinu, and Vlad Niculae. 2013. Temporal text classification for romanian novels set in the past. In *RANLP*.

Dalli, Angelo and Yorick Wilks. 2006. Automatic dating of documents and temporal text classification. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*. Association for Computational Linguistics.

Garcia-Fernandez, Anne, Anne-Laure Ligozat, Marco Dinarelli, and Delphine Bernhard. 2011. When was it written? automatically determining publication dates. In *String Processing and Information Retrieval*. Springer.

Graliński, Filip. 2013. Polish digital libraries as a text corpus. In *Proceedings of 6th Language & Technology Conference*, Poznań.

Guo, Siyuan, Trevor Edelblute, Bin Dai, Miao Chen, and Xiaozhong Liu. 2015. Toward enhanced metadata quality of large-scale digital libraries: Estimating volume time range. *iConference 2015 Proceedings*.

Jong, de FMG, Henning Rode, and Djoerd Hiemstra. 2005. Temporal language models for the disclosure of historical text. Royal Netherlands Academy of Arts and Sciences.

Kanhabua, Nattiya and Kjetil Nørvåg. 2009. Using temporal language models for document dating. In *Machine Learning and Knowledge Discovery in Databases*, pages 738–741. Springer.

Kumar, Abhimanu, Jason Baldridge, Matthew Lease, and Joydeep Ghosh. 2012. Dating texts without explicit temporal cues. *CoRR*, abs/1211.2290.

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

Wierzchoń, Piotr. 2009. Fotodokumentacja 3.0. *Language, Communication, Information*, 4:63–80.