Temporal text classification for Romanian novels set in the past

Alina Maria Ciobanu
Liviu P. Dinu
Octavia-Maria Şulea
Faculty of Mathematics and Computer Science
Center for Computational Linguistics
University of Bucharest
alinamaria.ciobanu@yahoo.com
ldinu@fmi.unibuc.ro
mary.octavia@gmail.com

Anca Dinu
Faculty of Foreign Languages
University of Bucharest
anca_d_dinu@yahoo.com

Vlad Niculae
University of Wolverhampton
vlad@vene.ro

Abstract

In this paper we look at a task in historical linguistics and the study of language development, namely that of identifying the time when a text was written. The novelty is that we evaluate our classifier and our selected features on literary texts having their action placed in the past and written so as to give off the impression of the respective epoch. We investigate several types of features and ultimately go with a very simple set of 10 features which very accurately classifies the texts based on the century they were actually written in. We use random forests to obtain high performance.

1 Introduction and Motivation

Determining the time when a document was written is a task not only with implications in cultural heritage but one which proves important to many other domains such as historical and literary criticism, diachronic linguistics, manuscript phylogeny and stemmatics, and the elaboration of critical theories about the author of the texts in question. A more practical, coarser grained approach is to classify according to the century in which a text was written, approach that we take in this paper.

Within many instances of this task, disputes between linguists and historians appear. For example, among the first texts written in Romania there are four religious texts, *Codicele Voroneţean*, *Psaltirea Scheiană*, *Psaltirea Voroneţeană* and *Psaltirea Hurmuzachi*, for which the dating is disputed between the 15th century (idea promoted by historians such as Nicolae Iorga) and the end of the 16th century (idea maintained by linguists such as Rosetti) (Tagliavini, 1972). Often times, the texts present characteristics of a translation, yet they are not original translations but modern copies of lost originals.

For Romanian, the 16th century represents the beginning of Romanian writing. In (Dimitrescu, 1994, p. 13) the author states that the modern Romanian vocabulary cannot be completely understood without a thorough study of the texts written in this period, which should be considered the source of the literary language used today. In the 17th century, some of the most important cultural events took place, such as the improvement of the education system and the establishing of several printing houses, and this led to a new development of the Romanian language (Dimitrescu, 1994, p.75). Then, in the 18th century, a diversification of the philological interests in Romania took place through writing the first Romanian-Latin bilingual lexicons, the draft of the first monolingual dictionary, the first Romanian grammar, and the earliest translations from French (Lupu, 1999, p. 29).

The transition to the Latin alphabet, which was a significant cultural achievement, is completed in the 19th century. The Cyrillic alphabet is maintained in Romanian writing until around 1850, afterwards being gradually replaced with the Latin alphabet (Dimitrescu, 1994, p. 270). The 19th century is marked by the conflict (and eventually the compromise) between etymologism and phonetism in Romanian orthography. In (Maiorescu, 1866) the author argues for applying the phonetic principle and several reforms are enforced for this purpose. In the 20th century, some variations regarding the usage of diacritics in Romanian orthography are noticed.

In this paper we approach an interesting version of the epoch disambiguation task, successfully disambiguating the century in which Romanian novels with the action set in the past and written so as to simulate the action's epoch *appear* have been written in. We used novels of Romanian writers Mihail Sadoveanu and Ştefan Agopian with the action developing in different time periods between the 16th to the 20th century. For training and evaluation we used a multitude of texts written in either one of the 5 centuries.

2 Related Work

The influence of the temporal effects in automatic document classification is analyzed in (Mourão et al., 2008; Salles et al., 2010). The authors state that a major challenge in building text classification models may be the change which occurs in the characteristics of the documents and their classes over time (Mourão et al., 2008). Therefore, in order to overcome the difficulties which arise in automatic classification when dealing with documents dating from different epochs, identifying and accounting for document characteristics changing over time (such as class frequency, relationships between terms and classes and the similarity among classes over time (Mourão et al., 2008)) is essential and can lead to a more accurate discrimination between classes.

Dalli and Wilks (2006) successfully apply a method for classification of texts and documents based on their predicted time of creation, proving that accounting for word frequencies and their variation over time is accurate. Kumar et al. (2012) argue as well for the capability of this method, of using words alone, to determine the epoch in which a text was written or the time period a document refers to.

The effectiveness of using models for individual partitions in a timeline with the purpose of predicting probabilities over the timeline for new documents is investigated in (Kumar et al., 2011; Kanhabua and Nørvåg, 2009). This approach, based on the divergence between the language model of the test document and those of the timeline partitions, was successfully employed in predicting publication dates and in searching for web pages and web documents.

In (de Jong et al., 2005) the authors raise the problem of access to historical collections of documents, which may be difficult due to the different historical and modern variants of the text, the less standardized spelling, words ambiguities and other language changes. Thus, the linking of current word forms with their historical equivalents and accurate dating of texts can help reduce the temporal effects in this regard.

Chambers (2012) states that applying timestamps to documents is, to some extent, similar to topic classification, focusing on choosing a time period instead of a topic, but also relating to temporal words and phrases which describe the time period to be determined and are often comprised in the investigated documents. Therefore, he argues for the inclusion of these temporal expressions into the learning system for automatic document dating and proposes such a model which obtains better results than previous generative models.

In (Mihalcea and Nastase, 2012) the authors introduced the task of identifying changes in word usage over time, disambiguating the epoch at word-level.

Recently, Stajner and Zampieri (2013) used stylistic features, such as lexical richness, to predict the century of historical Portuguese texts.

3 Approach

3.1 Datasets used

In order to investigate the diachronic changes and variations in the Romanian lexicon over time, we used a corpus containing texts ranging from the 16$^{\text{th}}$ to the 20$^{\text{th}}$ century, representing the five different stages in the evolution of the Romanian language, as discussed in the introduction. We used this corpus for feature selection, model training and evaluation, following the methodology described in Section 3.2.

We used this model to classify 20$^{\text{th}}$ century novels with action set in the past. The novels we used are shown in Table 2 along with the century in which the action takes place.

For preprocessing, we removed words that are irrelevant for our investigation, such as dates and numbers and non-textual annotations marked by non alphanumeric characters. We performed basic word segmentation, using whitespace and punctuation marks as delimiters and we lower-cased all words.

Century	Title
16	Codicele Todorescu Codicele Martian Coresi, Evanghelia cu învăţătură Coresi, Lucrul apostolesc Coresi, Psaltirea slavo-română Coresi, Târgul evangheliilor Coresi, Tetraevanghelul Manuscrisul de la Ieud Palia de la Orăştie Psaltirea Hurmuzaki
17	The Bible Miron Costin, Letopiseţul Ţării Moldovei Miron Costin, De neamul moldovenilor Grigore Ureche, Letopiseţul Ţării Moldovei Dosoftei, Viaţa si petreacerea sfinţilor Varlaam Motoc, Cazania Varlaam Motoc, Răspunsul împotriva Catehismului calvinesc
18	Antim Ivireanul, Opere Axinte Uricariul, Letopiseţul Ţării Românesti şi al Ţării Moldovei Ioan Canta, Letopiseţul Ţării Moldovei Dimitrie Cantemir, Istoria ieroglifică Dimitrie Eustatievici Braşoveanul, Gramatica românească Ion Neculce, O samă de cuvinte
19	Mihai Eminescu, Opere, v. IX Mihai Eminescu, Opere, v. X Mihai Eminescu, Opere, v. XI Mihai Eminescu, Opere, v. XII Mihai Eminescu, Opere, v. XIII
20	Eugen Barbu, Groapa Mircea Cartarescu, Orbitor Marin Preda, Cel mai iubit dintre pământeni

Table 1: Historical Romanian dataset, used for training and evaluation

3.2 Classifiers and features

The texts in the corpus (in Table 1) were split into chunks of 500 sentences in order to increase the number of sample entries and have a more robust evaluation. A quarter of the chunks were held out as a test set. On the training set, we experimented with several intuitive engineered features based on dictionaries, sentence length, stop word frequencies, and on word endings, but the most effective feature set turns out to be extremely simple.

We represented the texts using a simple bag-of-words model, applying *tf* re-weighting, and performed χ^2 feature selection. The ten best features turn out to classify both the training set and the test set without error. The classifier used is a random forest ensemble with 20 trees. The tree parameter `max_features`, the maximum number of features to consider in a split, is left at the default value of $\sqrt{d}$, where $d = 10$ is the number of features. There is no need for further search since the accuracy is perfect.

For comparison, a multinomial Naive Bayes classifier on the same feature set obtains 90.1% accuracy. To check whether the random forest actually learns to identify parts of the same document, we trained the same model using the document name as label. In this case, the accuracy with which the system assigned to a chunk the name of the document from which it was extracted was only 72.1%. However, the misclassifications happen mostly within century level. A chunk was assigned to a document from the correct century

Author	Title	Century
Agopian	Tobit	17
	Sara	17
	Tache de Catifea	19
	Manualul Întamplărilor	19
	Ziua Mâniei	20
Sadoveanu	Fraţii Jderi	16
	Neamul Şoimareştilor	17
	Baltagul	19
	Hanu Ancuţei	19
	Păuna Mică	20
	Nicoară Potcoavă	20

Table 2: Literary texts written in the 20th century used in our evaluation.

with 98.1% accuracy.

For understanding this phenomenon more clearly, we plotted the mean and standard deviation of each feature across the five centuries investigated in Figure 1.

The system was put together using the *scikit-learn* machine learning library for Python, version 0.14 (Pedregosa et al., 2011).

4 Results

On the held-out test set, our system obtains a perfect accuracy of 100%, as discussed in Section 3.2. We classified, using this system, the texts from Table 2. Because the interest is at document level, we did not split into chunks of 500 sentences, but because of *tf* normalization, this does not affect the results.

We examined the confidence (estimated class probability score) of the classification, which is the average of the probabilities given by the 20 trees in the randomized forest. Classification is very confident and places all texts in the century when they were actually written in, namely the 20th. From Agopian's texts, only *Ziua Mâniei* is not classified with 100% confidence, getting a 5% chance of being from the 19th century. Mihail Sadoveanu's text *Hanu Ancuței* is also given a 5% confidence for the 19th century, while *Nicoară Potcoavă* gets 5% for the 18th century, 10% for the 19th, leaving still a high confidence of 85% for the true class, 20th century.

5 Conclusions

Our results exhibit good performance. Despite the fact that the problem is simple, overfitting is effectively prevented by extreme feature selection and the features used promise to be useful in determining the period of some disputed writings from Romanian literature. It is interesting to see that the features contain pairs of old and new variants of the same word (*cari/ care*, *pre/ pe*), as well as only old variants of a word (*amu* for *acum*, *derept* for *drept*), and are mostly functional words.

It is possible that a justification similar to the one encountered in authorship attribution holds: authors can try to mimic the lexicon of the century where they are setting the action, and use rare, loaded words that set the frame for readers. But by counting very frequent functional words in temporal variations, such as the 10 best features extracted by our pipeline, we can find the signal of the contemporary language of the author, one difficult to fake.

In this paper we focused on temporal classification which can be a first step in many applications such as building a system for automatically translating between language stages. An interesting next step would be to extend the study at a lexical level and identify all forms of a word in order to create a map of its historical development, something also useful in the task mentioned above.

Acknowledgements

The contribution of the authors to this paper is equal. Research supported by a grant of the Romanian National Authority for Scientific Research, CNCS – UEFISCDI, project number PN-II-ID-PCE-2011-3-0959.

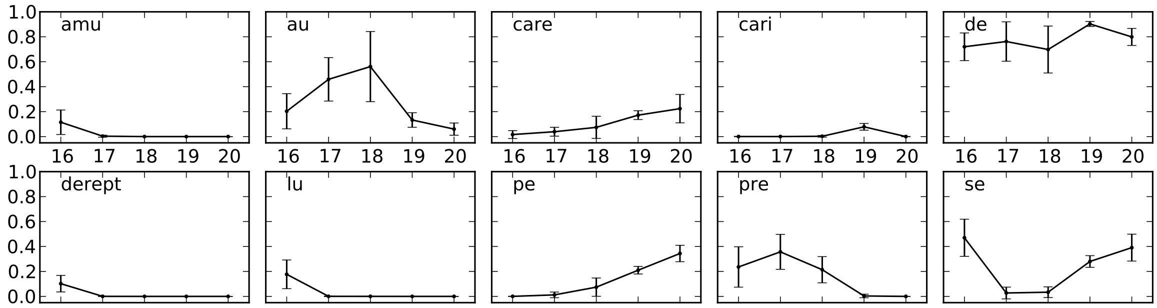

Figure 1: Mean and standard deviation of the keyword frequencies (y axis) for the 16-20 centuries (x axis). The translation of the feature words, from top to bottom and from left to right, are: old form of *now*, *(they) have*, modern form of *which*, old form of *which*, *of*, old form of *fair*, old form of *on*, modern form of *on*, reflexive form of the third person pronoun

References

Nathanael Chambers. 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*, ACL '12, pages 98–106, Stroudsburg, PA, USA. Association for Computational Linguistics.

Angelo Dalli 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, Sydney,*, pages 17—–22.

Franciska de Jong, Henning Rode, and Djoerd Hiemstra. 2005. Temporal language models for the disclosure of historical text. In *Humanities, computers and cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing*.

Florica Dimitrescu. 1994. *Dinamica lexicului românesc - ieri şi azi*. Editura Logos. In Romanian.

Nattiya Kanhabua and Kjetil Nørvåg. 2009. Using temporal language models for document dating. In *ECML/PKDD (2)*, pages 738–741.

Abhimanu Kumar, Matthew Lease, and Jason Baldridge. 2011. Supervised language modeling for temporal resolution of texts. In *CIKM*, pages 2069–2072.

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

Coman Lupu. 1999. *Lexicografia românească în procesul de occidentalizare latino-romanică a limbii române moderne*. Editura Logos. In Romanian.

Titu Maiorescu. 1866. Despre scrierea limbei rumăne. *Ediţiunea şi Imprimeria Societăţei Junimea*. In Romanian.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *ACL (2)*, pages 259–263. The Association for Computer Linguistics.

Fernando Mourão, Leonardo Rocha, Renata Araújo, Thierson Couto, Marcos Gonçalves, and Wagner Meira Jr. 2008. Understanding temporal aspects in document classification. In *WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 159–170.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, Oct.

Thiago Salles, Leonardo Rocha, Fernando Mourão, Gisele L. Pappa, Lucas Cunha, Marcos Gonçalves, and Wagner Meira Jr. 2010. Automatic document classification temporally robust. *Journal of Information and Data Management*, 1:199–211, June.

Sanja Stajner and Marcos Zampieri. 2013. Stylistic changes for temporal text classification. In *Proceedings of the 16th International Conference on Text Speech and Dialogue (TSD2013), Lecture Notes in Artificial Intelligence (LNAI 8082)*, pages 519–526, Pilsen, Czech Republic. Springer.

Carlo Tagliavini. 1972. *Le origini delle lingue neolatine*. Casa editrice Patron. In Italian.