Temporal classification for historical Romanian texts

Alina Maria Ciobanu
Liviu P. Dinu
Octavia-Maria Șulea
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 at border of natural language processing, historical linguistics and the study of language development, namely that of identifying the time when a text was written. We use machine learning classification using lexical, word ending and dictionary-based features, with linear support vector machines and random forests. We find that lexical features are the most helpful.

1 Introduction

Text dating, or determination of the time period when it was written, proves to be a useful component in NLP systems that can deal with such diachronistically dynamic inputs (Mourão et al., 2008). Besides this, the models that can perform such classification can shine light on less than obvious changes of certain features.

The knowledge captured in such systems can prove useful in transferring modern language resources and tools to historical domains (Meyer, 2011). Automatic translation systems between and across language stages, as in the corpus introduced by (Magaz, 2006), can benefit from the identification of feature variation over time.

In this paper we study the problem of supervised temporal text classification across genres and authors. The problem turns out to be solvable to a very high degree of accuracy.

2 Related Work

The influence of the temporal effects in automatic document classification is analyzed in (Mourão et al., 2008) and (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.

In (Dalli and Wilks, 2006) a method for classification of texts and documents based on their predicted time of creation is successfully applied, proving that accounting for word frequencies and their variation over time is accurate. In (Kumar et al., 2012) the authors 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 individuals partitions in a timeline with the purpose of predicting probabilities over the timeline for new documents is investigated in (Kumar et al., 2011) and (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.

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

3 Approach

3.1 Datasets used

In order to investigate the diachronic changes and variations in the Romanian lexicon over time, we used copora from five different stages in the evolution of the Romanian language, from the 16th to the 20th century. The 16th century represents the beginning of the 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 which led to the development of the Romanian language are the improvement of the education system and the establishing of several printing houses (Dimitrescu, 1994, p. 75). According to (Lupu, 1999, p. 29), in the 18th century a diversification of the philological interests in Romania takes 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. 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. To represent this period, we chose the journalism texts of the leading Romanian poet Mihai Eminescu. He had a crucial influence on the Romanian language and his contribution to modern Romanian development is highly appreciated. In the 20th century, some variations regarding the usage of diacritics in Romanian orthography are noticed.

Century	Corpus	Nwords	
		type	token
16	Codicele Todorescu	3,799	15,421
	Codicele Martian	394	920
	Coresi, Evanghelia cu învăţătură	10,361	184,260
	Coresi, Lucrul apostolesc	7,311	79,032
	Coresi, Psaltirea slavo-română	4,897	36,172
	Coresi, Târgul evangheliilor	6,670	84,002
	Coresi, Tetraevanghelul	3,876	36,988
	Manuscrisul de la Ieud	1,414	4,362
	Palia de la Orăştie	6,596	62,162
	Psaltirea Hurmuzaki	4,851	32,046
17	The Bible	15,437	179,639
	Miron Costin, Letopiseţul Ţării Moldovei	6,912	70,080
	Miron Costin, De neamul moldovenilor	5,499	31,438
	Grigore Ureche, Letopiseţul Ţării Moldovei	5,958	55,128
	Dosoftei, Viaţa si petreacerea sfinţilor	23,111	331,363
	Varlaam Motoc, Cazania	10,179	154,093
	Varlaam Motoc, Răspunsul împotriva Catehismului calvinesc	2,486	14,122
18	Antim Ivireanul, Opere	11,519	123,221
	Axinte Uricariul, Letopiseţul Ţării Românesti şi al Ţării Moldovei	16,814	147,564
	Ioan Canta, Letopiseţul Ţării Moldovei		
	Dimitrie Cantemir, Istoria ieroglifică	13,972	130,310
	Dimitrie Eustatievici Braşoveanul, Gramatica românească	5,859	45,621
	Ion Neculce, O samă de cuvinte	9,665	137,151
19	Mihai Eminescu, Opere, v. IX	27,641	227,964
	Mihai Eminescu, Opere, v. X	30,756	334,516
	Mihai Eminescu, Opere, v. XI	27,316	304,526
	Mihai Eminescu, Opere, v. XII	28,539	308,518
	Mihai Eminescu, Opere, v. XIII	26,242	258,234
20	Eugen Barbu, Groapa	14,461	124,729
	Mircea Cartarescu, Orbitor	35,486	306,541
	Marin Preda, Cel mai iubit dintre pământeni	28,503	388,278

Table 1: Romanian corpora: words

For preprocessing our corpora, we began by removing words that are irrelevant for our investigation, such as numbers. We handled word boundaries and lower-cased all words. We computed, for each text in our corpora, the number of words (type and token). The results are listed in Table 1. For identifying words from our corpora in dictionaries, we performed lemmatization. The information provided by the machine-readable dictionary *dexonline* [1] regarding inflected forms allowed us to identify lemmas (where no semantic or part-of-speech ambiguities occurred) and to further lookup the words in the dictionaries. In our investigations based on *dexonline* we decided to use the same approach as in (Mihalcea and Nastase, 2012) and to account only for unambiguous words. For example, the Romanian word *ai* is morphologically ambiguous, as we identified two corresponding lemmas: *avea* (verb, meaning *to have*) and *ai* (noun, meaning *garlic*). The word *amânare* is semantically ambiguous, having two different associated lemmas, both nouns: *amânar* (which means *flint*) and *amâna* (which means *to postpone*). We do not use the POS information di-

[1] http://dexonline.ro

rectly, but we use dictionary occurrence features only for unambiguous words.

The database of *dexonline* aggregates information from over 30 Romanian dictionaries from different periods, from 1929 to 2012, enabling us to investigate the diachronic evolution of the Romanian lexicon. We focused on four different subfeatures:

- words marked as obsolete in *dexonline* definitions (we searched for this tag in all dictionaries)
- words which occur in the dictionaries of archaisms (2 dictionaries)
- words which occur in the dictionaries published before 1975 (7 dictionaries)
- words which occur in the dictionaries published after 1975 (31 dictionaries)

As stated before, we used only unambiguous words with respect to the part of speech, in order to be able to uniquely identify lemmas and to extract the relevant information. The aggregated counts are presented in table 2.

Sub-feature		**16**	**17**	**18**	**19**	**20**
archaism	type	1,590	2,539	2,114	1,907	2,140
	token	5,652	84,804	56,807	120,257	62,035
obsolete	type	5,652	8,087	7,876	9,201	8,465
	token	172,367	259,367	199,899	466,489	279,654
< 1975	type	11,421	17,200	16,839	35,383	34,353
	token	311,981	464,187	337,026	885,605	512,156
> 1975	type	12,028	18,948	18,945	42,855	41,643
	token	323,114	480,857	356,869	943,708	541,258

Table 2: Romanian corpora: *dexonline* subfeatures

3.2 Classifiers and features

The texts in the corpus were split into chunks of 500 sentences in order to increase the number of sample entries and have a more robust evaluation. We evaluated all possible combinations of the four feature sets available:

- **lengths**: average sentence length in words, average word length in letters
- **stopwords**: frequency of the most common 50 words in all of the training set:

  de și în a la cu au no o să că se pe
  din s ca i lui am este fi l e dar pre ar
  vă le al după fost într când el dacă
  ne n ei sau sunt

Century	Precision	Recall	F1-score	texts
16	1.00	1.00	1.00	16
17	1.00	0.88	0.94	17
18	0.88	1.00	0.93	14
19	1.00	1.00	1.00	23
20	1.00	1.00	1.00	21
average/ total	0.98	0.98	0.98	91

Table 4: Random Forest test scores using all features and aggregating over 50 trees

- **endings**: frequency of all word suffixes of length up to three, that occur at least 5 times in the training set
- **dictionary**: proportion of words matching the *dexonline* filters described above

The system was put together using the *scikit-learn* machine learning library for Python (Pedregosa et al., 2011), which provides an implementation of linear support vector machines based on *liblinear* (Fan et al., 2008), an implementation of random forests using an optimised version of the CART algorithm.

4 Results

The hyperparameters (number of trees, in the random forest case, and C, for the SVM) were optimized using 3 fold cross-validation for each of the feature sets. For the best feature sets, denoted with an asterisk in table 3, the test results and hyperparameter settings are presented in tables 4 and 5.

The results show that the nonlinear nature of the random forest classifier is important when using feature sets so different in nature. However, a linear SVM can perform comparably, using only the most important features. The misclassifications that do occur are not between very distant centuries.

5 Conclusions

We presented two classification systems, a linear SVM one and a nonlinear random forest one, for solving the temporal text classification problem on Romanian texts. By far the most helpful features turn out to be lexical, with dictionary-based historical information less helpful than expected. This is probably due to inaccuracy and incompleteness of

lengths	stopwords	endings	dictionary	RF	SVM
False	False	False	False	25.38	25.38
False	False	False	True	86.58	79.87
False	False	True	False	98.51	95.16
False	False	True	True	97.76	97.02
False	True	False	False	98.51	96.27
False	True	False	True	98.51	94.78
False	True	True	False	98.88	*98.14
False	True	True	True	98.51	97.77
True	False	False	False	68.27	22.01
True	False	False	True	92.92	23.13
True	False	True	False	98.14	23.89
True	False	True	True	98.50	23.14
True	True	False	False	98.14	23.53
True	True	False	True	98.51	25.00
True	True	True	False	98.88	23.14
True	True	True	True	*99.25	22.75

Table 3: Cross-validation accuracies for different feature sets. The score presented is the best one over all of the hyperparameter settings, averaged over the folds.

Century	Precision	Recall	F1-score	texts
16	1.00	1.00	1.00	16
17	1.00	1.00	1.00	17
18	1.00	0.93	0.96	14
19	1.00	1.00	1.00	23
20	0.95	1.00	0.98	21
average/ total	0.99	0.99	0.99	91

Table 5: Linear SVC test scores using only stopwords and word endings for $C = 10^4$.

dictionary digitization, along with ambiguities that might need to be dealt with better.

We plan to further investigate feature importances and feature selection for this task to ensure that the classifiers do not actually fit authorship or genre latent variables.

Acknowledgements

The authors thank the anonymous reviewers for their helpful and constructive comments. 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.

References

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.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, June.

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.

Judit Martinez Magaz. 2006. Tradi imt (xx-xxi): Recent proposals for the alignment of a diachronic parallel corpus. *International Computer Archive of Modern and Medieval English Journal*, (30).

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

Roland Meyer. 2011. New wine in old wineskins? tagging old russian via annotation projection from modern translations. *Russian Linguistcs*.

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.