AMBRA: A Ranking Approach to Temporal Text Classification

Marcos Zampieri[1,2], Alina Maria Ciobanu[3], Vlad Niculae[4], Liviu P. Dinu[3]
Saarland University, Germany[1]
German Research Center for Artificial Intelligence (DFKI), Germany[2]
Center for Computational Linguistics, University of Bucharest, Romania[3]
Deptartment of Computer Science, Cornell University, USA[4]
marcos.zampieri@uni-saarland.de; alina.ciobanu@my.fmi.unibuc.ro;
vn66@cornell.edu; ldinu@fmi.unibuc.ro;

Abstract

This paper describes the AMBRA system, entered in the SemEval-2015 Task 7: 'Diachronic Text Evaluation' subtasks one and two, which consist of predicting the date when a text was originally written. The task is valuable for applications in digital humanities, information systems, and historical linguistics. The novelty of this shared task consists of incorporating label uncertainty by assigning an interval within which the document was written, rather than assigning a clear time marker to each training document. To deal with non-linear effects and variable degrees of uncertainty, we reduce the problem to pairwise comparisons of the form *is Document A older than Document B?*, and propose a non-parametric way to transform the ordinal output into time intervals.

1 Introduction

Temporal text classification consists of learning to automatically predict the publication date of documents, by using the information contained in their textual content. The task finds uses in fields as varied as digital humanities, where many texts have are unidentified or controversial publication dates, information retrieval (Dakka et al., 2012), where temporal constraints can improve relevance, and historical linguistics, where the interpretation of the learned models can confirm and reveal insights.

From a technical point of view, the task is usually tackled either as regression or, more commonly, as a single-label multi-class problem, with classes defined as time intervals such as months, years, decades or centuries. The regression approach assumes that precise timestamps are uniformly available for each document, which is suitable for cases of social media documents (Preotiuc-Pietro, 2014), but less suitable for documents surrounded by more uncertainty. Multi-class classification, on the other hand, suffers from a coarseness tradeoff: using coarser classes is less informative, and using finer classes reduces the number of training instances in each class, making the problem more difficult. Furthermore, with a multi-class formulation, the temporal relationship between classes is lost.

The 'Diachronic Text Evaluation' subtasks one and two from SemEval-2015 are formulated similarly to a multi-class problem, where each document is assigned to an interval such as 1976-1982. To accommodate such labels, we propose an approach based on pairwise comparisons. We train a classifier to learn which document out of a pair is older and which is newer. If two documents come from overlapping intervals, then their order cannot be determined with certainty, so the pair is not used in training. We use the property of linear models to extend a set of pairwise decisions into a ranking of test documents (Joachims, 2006).

While previous work uses a regression-based method to map the ranking back to actual timestamps, we propose a novel non-parametric method to choose the most likely interval. In light of this, our system is named AMBRA (Anachronism Modeling by Ranking). Our implementation is available under a permissive open-source license.[1]

[1] https://github.com/vene/ambra

2 Related Work

An important class of models for temporal classification employs prototype-based classification methods, using probabilistic language models and distances in distribution space to classify documents to the time period with the most similar language (de Jong et al., 2005; Kumar et al., 2011). Kanhabua and Nørvåg (2009) use temporal language models to assign timestamps to unlabeled documents.

An extension of such models for continuous time is proposed by Wang et al. (2008), who use Brownian motion as a model for topic change over time. This approach is simpler and faster than the discrete time version, but it cannot be directly applied to documents with different degrees of label uncertainty, such as interval labels.

Dalli and Wilks (2006) train a classifier to date texts within a time span of nine years. The method uses lexical features and it is aided by words whose frequencies increase at some point in time, most notably named entities. Abe and Tsumoto (2010) propose similarity metrics to categorise texts based on keywords calculated by indexes such as *tf-idf*. Garcia-Fernandez et al. (2011) explore different NLP techniques on a digitized collection of French texts published between 1801 and 1944. Style-related markers and features, including readability features, have been shown to reveal temporal information in English as well as Portuguese (Stamou, 2005; Štajner and Zampieri, 2013).

An intersecting research direction combines diatopic (regional) and diachronic variation for French journalistic texts (Grouin et al., 2010) and for the Dutch Folktale Database, which includes texts from different dialects and varieties of Dutch, as well as historical texts (Trieschnigg et al., 2012).

More recently, Ciobanu et al. (2013) propose supervised classification with unigram features with χ^2 feature selection on a collection of historical Romanian texts, noting that the informative features are words having changed form over time. Niculae et al. (2014) circumvent the limitations of supervised classification by posing the problem as ordinal regression with a learning-to-rank approach. They evaluate their method on datasets in English, Portuguese and Romanian. The superior flexibility of the ranking approach makes it a better fit for the problem formulation of the 'Diachronic Text Evaluation' task, motivating us to base our implementation on it.

A different, but related, problem is to model and understand how words usage and meaning change over time. Wijaya and Yeniterzi (2011) use the Google NGram corpus aiming to identify clusters of topics surrounding the word over time. Mihalcea and Nastase (2012) split the Google Books corpus into three wide epochs and introduce the task of *word epoch disambiguation*. Turning this problem around, Popescu and Strapparava (2013) use a similar approach to statistically characterize epochs by lexical and emotion features.

3 Methods

The 'Diachronic Text Evaluation' shared task consists of three subtasks (Popescu and Strapparava, 2015): classification of documents containing explicit references to time-specific persons or events (**T1**), classification of documents with time-specific language use (**T2**), and recognition of time-specific expressions (**T3**). The AMBRA system participated in T1 and T2.

3.1 Corpus

The training data released for the shared task consists of 323 documents for T1 and 4,202 documents for T2. Each document has a paragraph containing, on average, 71 tokens, along with a tag indicating when each text was written/published. The publication date of texts is indicated by time intervals at all three granularity levels: *fine-*, *medium-* and *coarse-grained* (e.g. `<textM yes="1695-1707">` for a text written between the years 1695 and 1707 in the medium-grained representation).

The shared task mentions no limitation regarding the use of external corpora. Nevertheless, to avoid thematic bias, we use only the corpora provided by the organizers under the assumption that the test and training sets are sampled from the same distribution.

The released test set consists of 267 instances for T1 and 1,041 instances for T2.

3.2 Algorithm and Features

We use a ranking approach by pairwise comparisons, previously proposed for temporal text modeling by Niculae et al. (2014) .

Learning. The model learns a linear function $g(x) = w \cdot x$ to preserve the temporal ordering of the texts, i.e. if document[2] x_i predates document x_j, which we will henceforth denote as $x_i \prec x_j$, then $g(x_i) < g(x_j)$. This step can be understood as *learning to rank* texts from older to newer. By making pairwise comparisons, the problem can be reduced to binary classification using a linear model.

A dataset annotated with intervals has the form $\mathcal{D} = \{(x, [y^{\text{first}}, y^{\text{last}})]\}$ where $y^{\text{first}} < y^{\text{last}}$ are the years between which document x was written. Document x_i can be said to predate document x_j only if its interval predates the other without overlap:

$$x_i \prec x_j \iff y_i^{\text{last}} < y_j^{\text{first}}.$$

This allows us to construct a dataset consisting only of correctly-ordered pairs:

$$\mathcal{D}_p = \{(x_i, x_j) : x_i \prec x_j\}.$$

This reduces to linear binary classification:

$$w \cdot x_i < w \cdot x_j \iff w \cdot (x_i - x_j) < 0.$$

We form a balanced training set by flipping the order of half of the pairs in D_p at random.

Prediction. Niculae et al. (2014), following Pedregosa et al. (2012), fit a monotonic function mapping from years to the space spanned by the learned linear model. In contrast, to better deal with the interval formulation, we propose a non-parametric memory-based approach. After training, we store:

$$D_{\text{scores}} = \{(z = w \cdot x, [y^{\text{first}}, y^{\text{last}}]\}.$$

When queried about when a previously unseen document x was written, we compute $z = w \cdot x$ and search for the k closest entries in D_{scores}, which we denote D^z_{scores}. For each candidate interval for the test document $[y^{\text{first}}, y^{\text{last}}]$ we compute its average distance to the intervals of the k nearest training documents $[y_i^{\text{first}}, y_i^{\text{last}}] \in D^z_{\text{scores}}$ where:

$$\text{dist}\,(y_a, y_b) = \left| \frac{y_a^{\text{last}} + y_a^{\text{first}}}{2} - \frac{y_b^{\text{last}} + y_b^{\text{first}}}{2} \right|.$$

The predicted interval is the one minimizing the average distance:

$$\hat{y} = \arg\min_{y \in \mathcal{Y}} \frac{1}{k} \sum_{y_i \in D^z_{\text{scores}}} \text{dist}(y, y_i).$$

Importantly, this approach allows for even more flexibility in interval labels than needed for the 'Diachronic Text Evaluation' task. While in the task all intervals (at a given granularity level) have the same size, our method can deal with intervals of various sizes,[3] half-lines $[-\infty, a]$ or $[a, \infty]$ for expressing only a lower or only an upper bound on the time of writing of a document, and even degenerate intervals $[a, a]$ for when the time is known exactly.

Features. AMBRA uses four types of features:

- Length meta-features (number of sentences, types, tokens);
- Stylistic (Average Word Length, Average Sentence Length, Lexical Density, Lexical Richness);[4]
- Grammatical (part-of-speech tag n-grams);
- Lexical (token n-grams).

We use χ^2 feature selection with classes defined as the $[50 \cdot n, 50 \cdot (n+1)]$ interval that overlaps the most with the true one. This coarse approach to feature selection has been shown to work well for temporal classification (Niculae et al., 2014).

4 Results

We perform 5-fold cross-validation over the training set to estimate the task-specific score. We fix the number of neighbours used for prediction to $k = 10$ after cross-validation using only number of tokens as feature. The model parameter space consists of the logistic regression's regularization parameter C, the minimum and maximum frequency thresholds for pruning too rare and too common features, n-gram range for tokens and for part-of-speech tags, and the number of features to keep after feature selection. We choose the best configuration after many

[2] We overload x_i to refer to the document itself as well as its representation as a feature vector.

[3] In our implementation, we set $\text{dist}(y_a, y_b)$ to 0 if the smaller interval is fully contained in the wider one.

[4] Lexical Density = unique tokens / total tokens; Lexical Richness = unique lemmas / total tokens.

		Task 1				Task 2			
Model	**Features**	**Fine**	**Medium**	**Coarse**	**MAE**	**Fine**	**Medium**	**Coarse**	**MAE**
Random	—	0.09	0.21	0.44	73.16	0.30	0.43	0.59	80.58
Ridge	lengths+style	0.15	0.32	0.52	67.94	0.33	0.59	0.77	54.77
AMBRA	lengths+style	0.12	0.26	0.48	74.67	0.38	0.58	0.75	57.00
AMBRA	full	0.17	0.38	0.55	63.24	0.60	0.77	0.87	31.74

Table 1: Evaluation of AMBRA and the baselines on the test data. We report the task-specific score (between 0 and 1, higher is better) for the three levels of granularity, as well as the mean absolute error (*MAE*, lower is better) for the fine level of granularity.

iterations of randomized search. We compare our ranking model to a ridge regression baseline, employing the document length meta-features and using the middle of the time intervals as target values. We also evaluate a random baseline where one of the candidate intervals is chosen with uniform probability. For evaluation, we use the task-specific metric defined by the organizers (Popescu and Strapparava, 2015), based on the number of interval divisions between the prediction and the right answer. For context, we also report the mean absolute error obtained by taking the center of the intervals as a point estimate of the year. Table 1 shows the performance of AMBRA and the baseline systems on the test documents. On T1, the full AMBRA system is the only to beat the random baseline in all metrics (95% confidence). On T2, where more data is available, AMBRA with length and style features outperforms ridge regression at fine granularity (95% confidence), and the full AMBRA system outperforms all others in all metrics (99% confidence).[5]

4.1 Most Informative Features

To better understand the performance of our method we analyze the most informative features selected by our best models. We use identical feature sets for both tasks, and while there are some common patterns, we observe important differences in the feature rankings, confirming that T1 and T2 are different enough in nature to warrant separate modeling.

Among the features useful for both tasks we find the length of a document in sentences highly predictive, with newer texts being longer. Also, the linguistic structure *determiner + singular proper noun* is predictive of older texts, while *adjective + singular noun* is predictive of newer texts. The decrease in use of the contraction *'d* is captured in both cases. From the lexical features, the word *letters* indicates older texts, corresponding to the decreasing use of mail as telecommunication became mainstream.

Words useful for T1 are more topic- and time-specific ones, such as *army*, *emperor*, *troops*, while the T2 model, possibly enabled by the larger amount of data, proves capable of detecting diachronic spelling variation (*publick* and *public* are both selected, with opposite signs), outdated words (*upon*), and more subtle stylistic changes such as the decrease in use of the Oxford comma (a comma followed by a conjunction at the end of a list).

5 Conclusion and Future Work

We propose a ranking-based method to handle interval prediction and account for uncertainty in temporal text classification. Our approach proved competitive in the Semeval-2015 'Diachronic Text Evaluation' subtasks one and two. The features we used are simplistic but effective. We expect performance to improve by including linguistic and etymology expertise in the feature engineering and selection process, as well as by including world knowledge through named entities and linked data.

Our model allows for arbitrary interval labels, which is more expressive and more realistic than the task formulation. We plan to refine collections of historical texts and tighten the annotation intervals wherever possible. Our implementation can be made more scalable by following the random sampling methodology of Sculley (2009).

[5] All significance results are based on 10000 boostrap iterations with bias correction.

Acknowledgments

The authors are thankful to Fabian Pedregosa for valuable discussion, to the anonymous reviewers for their helpful and constructive comments, and to the organizers for preparing and running the shared task. Liviu P. Dinu was supported by UEFISCDI, PNII-ID-PCE-2011-3-0959.

References

Hidenao Abe and Shusaku Tsumoto. 2010. Text categorization with considering temporal patterns of term usages. In *Proceedings of ICDM Workshops*.

Alina Maria Ciobanu, Liviu P. Dinu, Anca Dinu, and Vlad Niculae. 2013. Temporal classification for historical Romanian texts. In *Proceedings of LaTeCH*.

Wisam Dakka, Luis Gravano, and Panagiotis G. Ipeirotis. 2012. Answering general time-sensitive queries. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):220–235.

Angelo Dalli and Yorick Wilks. 2006. Automatic dating of documents and temporal text classification. In *Proceedings of ARTE*, Sidney, Australia.

Franciska de Jong, Henning Rode, and Djoerd Hiemstra. 2005. Temporal language models for the disclosure of historical text. In *Proceedings of AHC*.

Anne Garcia-Fernandez, Anne-Laure Ligozat, Marco Dinarelli, and Delphine Bernhard. 2011. When was it written? Automatically determining publication dates. In *Proceedings of SPIRE*.

Cyril Grouin, Dominic Forest, Lyne Da Sylva, Patrick Paroubek, and Pierre Zweigenbaum. 2010. Présentation et résultats du défi fouille de texte DEFT2010 où et quand un article de presse a-t-il été écrit? *Actes du sixième DÉfi Fouille de Textes*.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of KDD*.

Nattya Kanhabua and Kjetil Nørvåg. 2009. Using temporal language models for document dating. In *Proceedings of ECML/PKDD*.

Abhimanu Kumar, Matthew Lease, and Jason Baldridge. 2011. Supervised language modelling for temporal resolution of texts. In *Proceedings of CIKM*.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of ACL*.

Vlad Niculae, Marcos Zampieri, Liviu P. Dinu, and Alina Ciobanu. 2014. Temporal text ranking and automatic dating of texts. In *Proceedings of EACL*.

Fabian Pedregosa, Elodie Cauvet, Gael Varoquaux, Christophe Pallier, Bertrang Thirion, and Alexandre Gramfort. 2012. Learning to rank from medical imaging data. *CoRR*, abs/1207.3598.

Octavian Popescu and Carlo Strapparava. 2013. Behind the times: Detecting epoch changes using large corpora. In *Proceedings of IJCNLP*.

Octavian Popescu and Carlo Strapparava. 2015. Semeval-2015 task 7: Diachronic text evaluation. In *Proceedings of SemEval*.

Daniel Preotiuc-Pietro. 2014. *Temporal models of streaming social media data*. Ph.D. thesis, University of Sheffield.

D. Sculley. 2009. Large scale learning to rank. In *NIPS Workshop on Advances in Ranking*, pages 1–6.

Sanja Štajner and Marcos Zampieri. 2013. Stylistic changes for temporal text classification. In *Proceedings of TSD*.

Constantina Stamou. 2005. *Dating Victorians: An experimental approach to stylochronometry*. Ph.D. thesis, University of Bedfordshire.

Dolf Trieschnigg, Djoerd Hiemstra, Mariet Theune, Franciska de Jong, and Theo Meder. 2012. An exploration of language identification techniques for the dutch folktale database. In *Proceedings of LREC2012*.

Chong Wang, David Blei, and Heckerman David. 2008. Continuous time dynamic topic models. In *Proceedings of UAI*.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the Workshop on Detecting and Exploiting Cultural Diversity on the Social Web (DETECT)*.