DaDoEval @ EVALITA 2020: Same-Genre and Cross-Genre Dating of Historical Documents

Stefano Menini[*], Giovanni Moretti[**], Rachele Sprugnoli[**], Sara Tonelli[*]
[*]DH Research Group, Fondazione Bruno Kessler
Via Sommarive 18, 38123 Trento
[**]CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Agostino Gemelli 1, 20123 Milano
{menini,satonelli}@fbk.eu
{giovanni.moretti,rachele.sprugnoli}@unicatt.it

Abstract

English. In this paper we introduce the DaDoEval shared task at EVALITA 2020, aimed at automatically assigning temporal information to documents written in Italian. The evaluation exercise comprises three levels of temporal granularity, from coarse-grained to year-based, and includes two types of test sets, either having the same genre of the training set, or a different one. More specifically, DaDoEval deals with the corpus of Alcide De Gasperi's documents, providing both public documents and letters as test sets. Two systems participated in the competition, achieving results always above the baseline in all subtasks. As expected, coarse-grained classification into five periods is rather easy to perform automatically, while the year-based one is still an unsolved problem also due to the lack of enough training data for some years. Results showed also that, although De Gasperi's letters in our test set were written in standard Italian and in a style which was not too colloquial, cross-genre classification yields remarkably lower results than the same-genre setting.[1]

1 Introduction

In the context of EVALITA 2020 (Basile et al., 2020), we propose the task of assigning a temporal span to a document, i.e. recognising when a document was issued. The task has already been addressed in other languages, namely French, English, Polish, also in the framework of shared tasks, see for example the DÉfi Fouille de Textes (DEFT) 2010 and 2011 challenges (Grouin et al., 2010; Grouin et al., 2011), the SemEval-2015 task on Diachronic Text Evaluation (Popescu and Strapparava, 2015) and the RetroC challenge (Graliński et al., 2017). This task is relevant because it can play a role in document retrieval, summarisation, event detection, etc. It is also an important task per se, since it can be used to process large archival collections. In particular, when some documents in a collection have not been dated, supervised approaches could be applied to learn from the documents with a date which time span can be assigned to those who are not provided with temporal metadata. Along this line, we proposed our task taking Alcide De Gasperi's corpus of public documents (Tonelli et al., 2019) as a use case. To our knowledge, this task for Italian has never been proposed before to the NLP community, which means that all participating systems have been built from scratch.

All information related to the task, the official scorer and the training, test and gold data are available on the task website https://dhfbk.github.io/DaDoEval/.

2 Task Description

The goal of the DaDoEval shared task is to foster the development of systems able to automatically assign temporal information to unseen documents with different granularity. Therefore, we foresee three types of temporal spans, from coarse-grained to year-based, corresponding to different classification difficulty. Furthermore, we want to assess the impact of out-of-domain data on classification quality. We therefore propose the six following subtasks:

1a **Coarse-grained classification on same-genre data**: participants are asked to assign each document in the test set to one of the main time periods that historians have identi-

A	B	C	D	E
Habsburg years	Beginning of political activity	Internal exile	From fascism to the Italian Republic	Building the Italian Republic
1901-1918	1919-1926	1927-1942	1943-1947	1948-1954

Table 1: Time periods for the coarse-grained tasks.

fied in De Gasperi's life, reported in Table 1. Each document in the training set is labeled with one of the five periods and test data are of the same genre of the training data, both taken from the corpus of De Gasperi's public documents (Tonelli et al., 2019).

1b **Coarse-grained classification on cross-genre data**: participants are asked to assign each document in the test set to one of the main time periods that historians have identified in De Gasperi's life, reported in Table 1. Each document in the training set is labeled with one of the five periods and taken from the corpus of De Gasperi's public documents, while the test set contains letters from De Gasperi's correspondence (Tonelli et al., 2020).

2a **Fine-grained classification on same-genre data**: participants are asked to assign each document in the test set to one temporal slice of 5 years. Each document in the training set is labeled with a temporal slice and test data are of the same genre of the training data, both taken from De Gasperi's public documents.

2b **Fine-grained classification on cross-genre data**: participants are asked to assign each document in the test set to one temporal slice of 5 years. Each document in the training set is labeled with a temporal slice and test data are extracted from De Gasperi's correspondence.

3a **Year-based classification on same-genre data**: participants are asked to assign each document in the test set to its exact year of publication. Each document in the training set is labeled with the year of publication and test data are of the same genre of the training data, both taken from De Gasperi's public documents.

3b **Year-based classification on cross-genre data**: participants are asked to assign each document in the test set to its exact year of publication. Each document in the training set is labeled with the year of publication and test data are extracted from De Gasperi's correspondence.

Subtask 1 is the easiest task of the challenge, since the five time periods were defined by history scholars based also on the different roles and events involving De Gasperi during his career. We expect therefore that the documents grouped together for each time period present a high degree of similarity concerning topics, mentioned people and events. Also different document types should vary over time, with more news articles dated between 1901 and 1918, when De Gasperi worked as a journalist, and more telegrams written towards the end of his career, when De Gasperi was Minister of Foreign Affairs.

Subtask 2 includes 11 classes, each comprising 5 years. In this case, however, the division is arbitrary and purely based on the document date, therefore documents in the same class do not necessarily have anything in common concerning the topic, De Gasperi's role, etc. Finally, subtask 3 is the most challenging one, also because for some years only few training examples were available. More details on the document distribution in the training set are reported in Section 3.

The aforementioned subtasks can be addressed in several ways. For example, researchers interested in historical content analysis can infer temporal information by looking at persons, places and time expressions, possibly integrating linking techniques. For those interested in studying semantic shifts, a purely lexical analysis may highlight changes in the lexical choices made by De Gasperi over time and give hints for document dating (Kulkarni et al., 2018). Also deep learning techniques, which proved effective on larger English corpora for document dating, could be tested (Vashishth et al., 2018). As an alternative, the subtasks could be addressed using document similarity techniques, so to assess to which training documents those in the test set are most similar, as-

suming that similar documents have been written in the same years.

3 Dataset

The corpus of De Gasperi's public documents contains 2,759 documents, manually tagged with a date, written by De Gasperi and issued between 1901 and 1954. All the documents have been written by the same person, thus removing the effects that different author styles can have on the dating process. Since we proposed a supervised task, the corpus was split into a training and a test set following an 80:20 ratio, thus having 2,210 documents for training and the remaining 549 for testing.

In addition to the in-domain test set, we also provide a cross-genre out-of-domain test set of 100 private letters, written by De Gasperi in the same time span of the corpus of public documents within the Epistolario project[2]. This out-of-domain test set allowed DaDoEval organisers to evaluate the robustness of the proposed approaches, and measure how the specific characteristics of correspondence affect the dating process.

We report in Table 3 the document distribution in the training and test set for the coarse- and the fine-grained subtasks. In general, the classes are not well-balanced, with some periods having only few training documents. For example, in the fine-grained subtask the span 1926 – 1930 has only 16 documents vs. 599 documents belonging to the period 1946 – 1950.

In Figure 1 and 2 we show also the year-based distribution of documents in the training and in the test set. While the same-genre distribution is similar, the letters in the test set (red line in the graph) are more homogeneous, with no year-based peaks like for public documents. On the contrary, some years that are barely represented in the training set (for example 1927) present several instances in the cross-genre test set, making classification particularly challenging.

For both corpora, there are no privacy issues and the documents can be made freely to task participants.

4 Evaluation Procedure and Baseline

Each participating team is allowed to submit two runs for each subtask. The evaluation is performed by computing class-based Precision, Recall and F1, and then the macro-averaged F1, upon which the final ranking is based. The task scorer is available on the task website[3].

As a baseline, we adopt for all tasks the same Logistic Regression configuration. As features to represent the document content, we calculate tf-idf for each term (unigram) in the dataset, without removing stopwords or performing any preprocessing on the text. For computing tf-idf and training the Logistic Regression classifier we rely on the scikit-learn library (Pedregosa et al., 2011).

		Same-genre		Cross-genre
		Train	Test	Test
Coarse-grained	class1	572	140	20
	class2	342	109	20
	class3	150	37	20
	class4	514	98	20
	class5	632	165	20
Fine-grained	1901-1905	85	21	3
	1906-1910	256	65	6
	1911-1915	211	48	5
	1916-1920	109	42	11
	1921-1925	246	73	12
	1926-1930	16	2	10
	1931-1935	76	22	4
	1936-1940	62	13	8
	1941-1945	191	36	15
	1946-1950	599	129	16
	1951-1955	399	98	10

Table 2: Document distribution for the coarse-grained and the fine-grained subtasks.

5 Participants and Results

Eighteen teams registered to participate, but only two actually submitted the results for the evaluation for a total of 16 runs. Both participants come from the academia: one from Italy (University of Pisa) and one from Germany (University of Tübingen). A short description of each system follows:

matteo-brv (University of Tübingen) participated only in subtask 1 and 2 with two runs for each subtask (Brivio, 2020). Both subtasks have been treated as classification problems and modeled with a linear Support Vector Machine multi-class

[2] https://www.epistolariodegasperi.it/

[3] https://github.com/dhfbk/DaDoEval/blob/master/DaDoEval_Eval.py

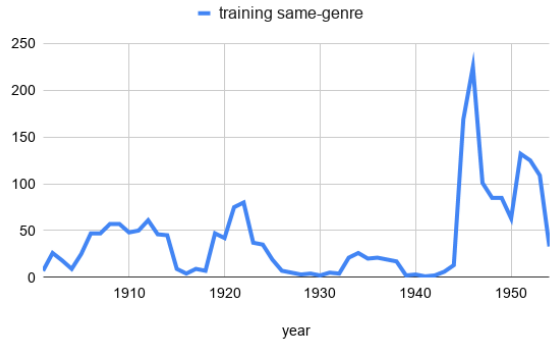

Figure 1: Per year document distribution in the training set.

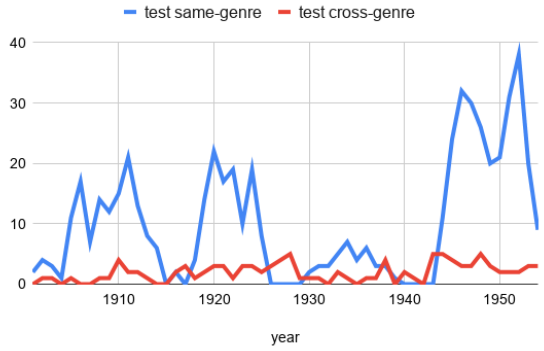

Figure 2: Per year document distribution in the test set.

Same-genre								
subtask 1a			subtask 2a			subtask 3a		
TEAM	#RUN	MACRO F1	TEAM	#RUN	MACRO F1	TEAM	#RUN	MACRO F1
matteo-brv	1	0.934	rmassidda	2	0.638	rmassidda	2	0.274
matteo-brv	2	0.934	rmassidda	1	0.579	rmassidda	1	0.256
rmassidda	1	0.858	BASELINE		0.485	BASELINE		0.126
rmassidda	2	0.855						
BASELINE		0.827						

Cross-genre								
subtask 1b			subtask 2b			subtask 3b		
TEAM	#RUN	MACRO F1	TEAM	#RUN	MACRO F1	TEAM	#RUN	MACRO F1
matteo-brv	1	0.413	rmassidda	2	0.177	rmassidda	1	0.074
matteo-brv	2	0.413	BASELINE		0.171	rmassidda	2	0.035
rmassidda	2	0.392	rmassidda	1	0.158	BASELINE		0.02
BASELINE		0.368						
rmassidda	1	0.366						

Table 3: Results of six subtasks in terms of macro-average F1.

classifier, implemented through the scikit-learn library (Pedregosa et al., 2011). The model was trained on a set of style-based features: TF-IDF weighted character and word n-grams, and number of word tokens per document. Features have been extracted without any form of data set pre-processing. N-gram size has been determined empirically and found to yield the best results in a range of 3 to 5 and 1 to 2 for character and word n-grams, respectively. On the other hand, TF-IDF parameters and model parameters were tuned using a 5-fold cross validation Bayesian optimization strategy, an algorithm implemented in the Scikit-Optimize library[4].

rmassidda (University of Pisa) participated in all subtasks with 2 runs for each of them (Massidda, 2020). Two representations are generated for each document with no fine-tuning: (i) a sequence of sentence embeddings using Sentence-BERT (Reimers and Gurevych, 2019), and (ii) a bag-of-entities obtained using the spaCY Named Entity Recognition system[5]. Since the performance obtained on a validation set showed that the first representation yields better results on the coarse-grained task, while the bag-of-entities performed better on the fine- and year-based tasks, the two representations are combined in an architecture where the sentence embeddings are fed to a transformer block containing a multi-headed self-attention layer. Its output is then averaged and concatenated with the bag-of-entities representation of the document before being fed to a multi-layer neural network. The output of each

[4] https://scikit-optimize.github.io/stable/

[5] https://github.com/explosion/spacy-models

layer of this network is also fed to a dedicated neural network that produces the output of each subtask.

6 Discussion

6.1 System comparison

The two submitted systems are based upon different paradigms: **matteo-brv** relies on an SVM-based classifier with simple linguistic features, while **massidda** uses recent transformer-based models and neural networks. Despite being more computationally intensive and complex, the second approach yields a lower performance than the first one. The difference in performance, however, is smaller in the cross-genre subtask (0.02 F1) than in the same-genre one (0.07 F1). As a comparison, we show in Fig. 3 the average F1 obtained by each participant's best run for the five classes (i.e. time periods) in the same-genre coarse-grained task. The results across the five classes are rather balanced and do not reflect the number of training examples for each class (see Table 3). Indeed, Class 3 (from 1927 to 1942) has the least number of training documents but both systems achieve the best results. This probably depends on the fact that in those years De Gasperi does not participate in public life and has no political role, therefore the tone, topics and mentioned people are probably different from those in the rest of the document collection, therefore they are easily identifiable.

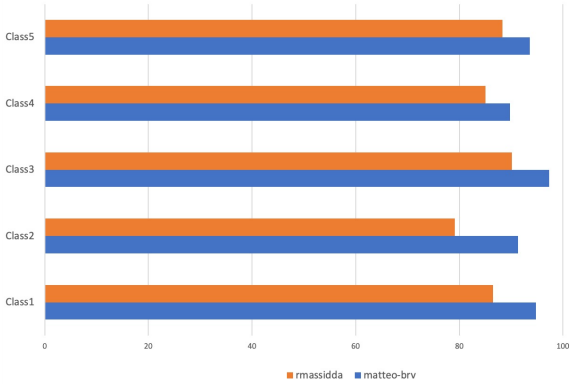

Figure 3: Comparison of participating systems on same-genre coarse-grained task

In Figure 4 we report the same comparison but in the cross-genre coarse-grained task. In this case, the two systems show a completely different behaviour, obtaining the worse results on Class 3. Furthermore, no system achieves the best result on all classes, like for the same-genre task. Interestingly, on Class 2 and 3, containing the least training documents, the neural approach by rmassidda clearly outperforms the SVM-based one.

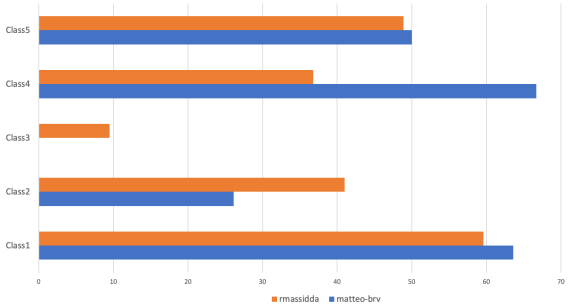

Figure 4: Comparison of participating systems on cross-genre coarse-grained task

Overall, there are huge performance differences with different classification granularity: while the coarse-grained subtask on same-genre data achieves a macro F1 above 0.82 even with a simple logistic regression baseline, performance drops dramatically with the fine-grained classification, and in the year-based task every presented approach yields insufficient results for any practical application. The presence of 55 classes (i.e. years) as well as an unbalanced distribution of training instances in the different classes make it indeed very difficult to build a robust supervised system.

After the competition deadline, matteo-brv submitted with the same SVM-based configuration the runs for subtasks 2 and 3, which were missing in the original submission. If regularly submitted to the competition, the system performance would be top-ranked with 0.702 in subtask 2a, 0.403 in subtask 3a, 0.240 on subtask 2b and 0.086 on subtask 3b. This confirms that, when dealing with middle-sized datasets, non-neural approaches can still be the best option, beside being easier to tune and less computationally intensive than neural classifiers.

6.2 Dataset comparison

In order to understand the impact of genre on classification performance, we randomly select 20 documents for each time period in the same-genre test set so to obtain a subcorpus similar in size (100 documents) and distribution as the cross-genre test set. Then, we process both corpora by running the Tint NLP Suite (Aprosio and Moretti, 2018), using in particular the modules computing complexity and readability indices.

From a lexical point of view, the two test sets do not differ much. For instance, type-token ratio is 0.81 in the same-genre subcorpus and 0.79 in the cross-genre one. In both cases, the value is rather high, confirming the careful selection of terms and expressions performed by De Gasperi, who was well-known for formal, sometimes archaic use of the language. This is evident also in the letters, even if they concerned people and events from his private sphere. Also the lexical density, i.e. the proportion between content words and the total number of words, is very similar, being 0.58 in same-genre subcorpus and 0.59 in the cross-genre one. Also in this case, the higher the value, the more 'conceptually dense' the text is, requiring more cognitive effort to read and understand the document content.

Although from a lexical point of view the two subcorpora are aligned, we observe a difference from the syntactic point of view. Indeed, while the average sentence length in the same-genre subcorpus is 21 tokens, it is 13 in the letters. This difference is confirmed also by the Gulpease score (Lucisano and Piemontese, 1988), which is the standard readability metric for Italian taking into account word and sentence length as a proxy for complexity. Gulpease is 61 for the letters and 50 for the same-genre subcorpus, corresponding to a higher readability for the former (the higher, the easier to read). Overall, this analysis shows that the more informal style usually associated with letters is expressed by De Gasperi through the use of simpler syntactic structures rather than through a simpler vocabulary. Also, classification approaches that rely on sentence-based units, for example sentence embeddings, may perform worse when the sentence characteristics are very different in the training and the test set.

If we consider semantic information, we observe also in this case some differences. For instance, the use of named entities is less frequent in letters than in the same-genre test set (0.44 avg. NER per sentence vs. 0.58). This holds for all the NER types considered, from persons (0.19 per sentence vs. 0.21) to locations (0.14 vs. 0.21). This again may affect the performance of systems using NER-based analysis like bag-of-entities, when the use of NER varies a lot between the training and the test set.

7 Conclusions

In this paper we have presented the DaDoEval task, which has been proposed for the first time at EVALITA 2020, with the goal to automatically date Italian documents. The task includes three different classification granularities, from five broad time spans to fifty-five years. Two subtasks are also foreseen, i.e. same-genre and cross-genre classification. The corpus used is the collection of De Gasperi's public documents, plus 100 letters being the test set for the cross-genre task.

Two systems have participated in the DaDoEval evaluation exercise, but only for the coarse-grained setting. In the other subtasks, there has been only one participant. A comparison between the two approaches has showed that a classifier based on SVM has consistently achieved better results than a neural one even if using a much simpler architecture. We also observed that cross-genre classification is still problematic, as is fine-grained classification. In order to have a better understanding of fine-grained classification, and provide more insightful system comparisons, it would be interesting to modify the scorer so to take into account how close misclassified examples are from the correct year or time period. This would provide a partial recognition to wrong instances when the assigned date is not far from the correct one.

The datasets and the scorer have been made available to the research community through the DaDoEval website, so that researchers will be able to deal with this task in the future, which is far from being solved.

Acknowledgements

We thank the President of the National Edition of De Gasperi's Letters Giuseppe Tognon and Stefano Malfatti for giving us access to the letters used in cross-genre classification task.

References

Alessio Palmero Aprosio and Giovanni Moretti. 2018. Tint 2.0: an All-inclusive Suite for NLP in Italian. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Matteo Brivio. 2020. matteo-brv @ DaDoEval: An SVM-based Approach for Automatic Document Dating. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Filip Graliński, Rafał Jaworski, Łukasz Borchmann, and Piotr Wierzchoń. 2017. The RetroC challenge: how to guess the publication year of a text? In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, pages 29–34.

Cyril Grouin, Dominic Forest, Patrick Paroubek, and Pierre Zweigenbaum. 2010. Présentation et résultats du défi fouille de texte deft2011 quand un article de presse a t-il été écrit? à quel article scientifique correspond ce résumé? In *Actes du septième DÉfi Fouille de Textes*.

Cyril Grouin, Dominic Forest, Patrick Paroubek, and Pierre Zweigenbaum. 2011. Présentation et résultats du défi fouille de texte deft2011 quand un article de presse a t-il été écrit? à quel article scientifique correspond ce résumé? In *Actes du septième DÉfi Fouille de Textes*.

Vivek Kulkarni, Yingtao Tian, Parth Dandiwala, and Steven Skiena. 2018. Simple neologism based domain independent models to predict year of authorship. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 202–212.

Pietro Lucisano and Maria Emanuela Piemontese. 1988. Gulpease. Una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e Città*, 3:57–68.

Riccardo Massidda. 2020. rmassidda @ DaDoEval: Document Dating Using Sentence Embeddings at EVALITA 2020. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

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.

Octavian Popescu and Carlo Strapparava. 2015. Semeval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 870–878.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983.

Sara Tonelli, Rachele Sprugnoli, and Giovanni Moretti. 2019. Prendo la Parola in Questo Consesso Mondiale: A Multi-Genre 20th Century Corpus in the Political Domain. In *In Proceedings of CLiC-it 2019*.

Sara Tonelli, Rachele Sprugnoli, Moretti Giovanni, Malfatti Stefano, and Odorizzi Marco. 2020. Epistolario De Gasperi: National Edition of De Gasperi's Letters in Digital Format. In *IX Convegno Annuale AIUCD*, pages 253–259. Alma Mater Digital Library.

Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2018. Dating documents using graph convolution networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1605–1615.