SemEval 2015, Task 7: Diachronic Text Evaluation

Octavian Popescu
IBM Research
Yorktown, NY 10598, USA
o.popescu@us.ibm.com

Carlo Strapparava
FBK
Povo, TN 38123, Italy
strappa@fbk.eu

Abstract

In this paper we describe a novel task, namely the Diachronic Text Evaluation task. A corpus of snippets which contain relevant information for the time when the text was created is extracted from a large collection of newspapers published between 1700 and 2010. The task, subdivided in three subtasks, requires the automatic system to identify the time interval when the piece of news was written. The subtasks concern specific type of information that might be available in news. The intervals come in three grades: fine, medium and coarse according to their length. The systems participating in the tasks have proved that this a doable task with very interesting possible continuations.

1 Introduction

Language changes over the time, even over relatively small periods. For example, as the main intent of publishing newspapers is to disseminate information to the population of a whole country, there is an objective pressure to impose a standard and to smooth over the dialectical differences. However, since the late 1600s, each generation has read pieces of news containing new words, borrowed or invented, exhibiting new drifts in the meanings of old words, printed with different spelling etc.

The examples (1), (2), (3) and (4) below exhibit a series of features which are useful to pin point the year when the respective piece of news was created. Well known global events, sense superseding, specific spelling and new vocabulary entry words are all time relevant features. At a deeper level of analysis, time is revealed also by the mentions of named entities, such as *Security Council*, the topic and the linguistic genre are also relevant features.

1. Dictator Saddam Hussein ordered his troops to march into Kuwait. After the invasion is condemned by the UN Security Council, the US has forged a coalition with allies. Today American troops are sent to Saudi Arabia in Operation Desert Shield, protecting Saudi Arabia from possible attack. **circa 1990**

2. We have cabled the English house from which we get it and expect a reply to-morrow. **circa 1900**

3. Occasional selfies are acceptable, but uploading a new picture of yourself every day is not necessary. **circa 2014**

4. ...The House of Samuel Sandbroke was brokt and several Pistols discharged ...Her Majesty, for the better Discovery of the Offenders, is pleased to promise Her most Gracious Pardon for the said Crime. **circa 1705**

While for humans it is relatively easy to notice the language differences between two texts, and even to be accurate in determining the period when a piece of news was written, for computational systems this task is challenging. On the other hand, with the availability of large time-tagged corpora, a computational system can perform various analyses and extract correlations that are impossible for humans to know beforehand or acquire through manual inspection of the information scattered over huge collections of texts.

We propose to tackle the task of automatically identifying the time period when a piece of news was written. We provide a corpus of fragments of pieces

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 870–878,
Denver, Colorado, June 4-5, 2015.

of news, for both training and testing. The interesting question is whether it is possible to automatically determine the period when a text was written. To this end, we have devised a SemEval 2015 task, called *Diachronic Text Evaluation*, hence DTE task. For this task, all aspects of language change may be taken into account and systems of various levels of analysis can be developed. The systems could benefit for a training corpus and are evaluated against a gold standard.

Organizing a diachronic task has proven to be a difficult one and we made decisions regarding what type of pieces of news are selected, what type of information they contain and how the evaluation could be carried out. In a nutshell, we have selected pieces of news of variable length ranging from ten to a couple of hundred words, and we have made a differentiation between pieces of news that mention famous named entities and those which do not. Our definition of famous is associated with the possibility of finding information about the respective named entities in external resources, such as Wikipedia. Consequently, we proposed two subtasks according to the difference above. For both tasks, the system has to guess the correct time interval in which the text was created. The intervals come in three types: fine, medium and coarse, according to their length. The third and last subtasks regard the phrases that carry time information and the systems only have to decide if a certain phrase in a given context is time relevant, and not to assign a precise time interval to the text.

The systems could use any type of algorithm to analyze the text and find the time relevant information. In fact, the main goal of the task was to identify fragments of text which by themselves, or in conjunction with a publicly available external resource, are time relevant. As such, the task is a systematic investigation into the actual capacity of NLP to combine both textual and meta-textual information in order to place a piece of text into a larger, temporal, context.

To the best of our knowledge, the present task is one of the very first systematic investigation in diachronic corpora with a focus on the textual and meta-textual features that are time relevant. We believe that systems for finding diachronic information for pieces of text are very interesting from both theoretical and practical point of view. Socio and historical linguistics are both based on the analyses of specific linguistics variability in a certain epoch, location, social class etc. The statistical methods are able to discover correlations and linguistic provable evidence of language change at all levels: morphological, syntactical, semantic and discourse. It would be physically impossible for a human, or a team of humans for what it matters, to analyze and corroborate the data from hundreds of gigabytes of data and find all the relevant differences. Looking at the distribution of words across timeline, salient periods, with statistically non-random behavior, can be automatically inferred (Popescu and Strapparava, 2013). The structure of such periods, or epochs, are by far more complex than what it could be manually performed. From a practical point of view, diachronic systems have a wide range of applications from emergent fields such as computational forensics, computational journalism to more traditional tasks, such as discourse similarity, sense shifting, readability and narrative frameworks, etc.

The paper is organized as follow: in the next section we review the relevant literature. In Section 3 we present the main motivation for the DTE task and the three subtasks with their specific corpora. In Section 4 we present the data format and the evaluation method together with a simple baseline. In Section 5 we discuss the main properties of the submitted systems and their results. The paper ends with a substantial section on conclusion and main future research direction in DTE.

2 Related Work

The availability of large time annotated corpora like Google N-gram open the perspective of a new field of the research which focuses on the distribution of the linguistics elements in certain periods. (Popescu and Strapparava, 2014; Popescu and Strapparava, 2013) showed how such corpora can be used to infer transition periods between epoch with specific characteristics. A ground breaking paper, (Niculae et al., 2014) focuses on historical documents in three languages, English, Portuguese and Romanian. The paper shows how statistical method can be used to predict the date when the documents have been created. The similarity of the ideas in the present task and their paper, although developed in completely

autonomy, prove that there is indeed a major interest in building diachronic systems and that the time is high for this task. We believe that there is a lot to do in this emergent field.

3 Task Description

In this section we present the main motivations for a diachronic task and in particular, we focus on how these motivations have influenced the choice in the present task. Let us start from the example (1)-(4) presented in Section 1.

We can observe that the choice of words, the morphology and word particular meaning, are an important part of time detection. Words like *brokt*, *selfie*, spellings like *to-morrow* or a sense like the one of the verb *cable* in (2) are used only within a certain period. Also, the topics are time specific and the reader may not even need to consult other sources in order to realize that an *American war in Saudi Arabia* and *Her Majesty pardon for a domestic incident* cannot possibly happen in the same period, as much as *telegraphing* and *uploading selfies* cannot either. Any of these clues seems to be a strong clue, but it would have been difficult to consider them before seeing this particular set of sentences. Intuitively, if one would read another set of sentences, some other clues, equally strong, are found. It makes sense to ask ourselves: How many such clues exist? Can such clues be systematically found and consistently organized? A human investigation of large corpora is hopeless, as billions of sentences must be inspected.

3.1 From News Corpora to Diachronic Data and Tasks

To answer this question we may want to link the linguistic information to the timeline. A big quantity of data, chronologically ordered, allows accurate statistical statements regarding the covariance between the frequencies of two or more terms over a certain period of time. By discovering significant statistical changes in word usage behavior, it is possible to define epoch boundaries. Inside these epochs the news are written in a rather uniform way. However, small changes as well as reference to famous historical events may lead to the formation of sub epochs.

Clearly, the mentioning of specific historical events makes it much easier for a diachronic system. The system must be able to consult an external resource such as Wikipedia, in order to assign a time stamp to the extracted entities. However, an extra analysis is required in order to make sure that the text does not refer to the respective historical event as past experience. On the other hand, surface features, such as spelling, reference to institutions that are specific to a epoch, or the usage of words in specific context, can be used to infer a time interval within which the text was written. Generally, this interval is much larger when compared to the time stamp assigned to the historical events and unless the system is provided with a crystal globe, no more accurate predictions can be made. It becomes clear that one needs to differentiate between the two types of information discussed above. And, also, that different precision is to be expected between these two subtasks. Let us call subtask 1 the diachronic task which considers pieces of news in which specific historical events, named entities etc. are clearly mentioned and let us call subtask 2 the diachronic task in which such information is missing, but in which there is enough surface information to assign a time interval, at least for a human. We present and discuss below a few typical examples for each of the subtasks mentioned here.

Task 1

5. At the Court at St.James's, the 29th Day of March, and1744 Present, the King's most excellent Majesty in Council. His Majesty's Declaration of War against the French King.

6. The Troubles which broke out in Germany on Account of the Succession of the late Emperor Charles the Sixth, having been begun, and carried on, by the Instigation, Assistance, and Support of the French King

7. By 1971 about one-third of Edison's electric output will be generated with nuclear capacity,

8. 1935 Ford V-8 Tudor Sedan Only an year old. not a flaw in it anywhere.

In example (5) the date is clearly indicated and the phrase *war against the French King* anchors the text very precisely in time. The mention of *the late Emperor Charles the Sixth* in example (6) pinpoints the time very precisely. The epoch is indicated in example (7) as *nuclear capacity* cannot possible happen before mid sixties. The last example, (8),

requires a subtraction of the dates expressed via temporal phrase, *1935* and *one year old* respectively. To sum up, task 1 requires systems to work with temporal expressions, name entity recognition and external resources, such as Wikipedia.

Task 2

9. By Letters from the Frontiers there is Advice, that the French Intendant has given Orders for tracing out a Camp near Givet for 10000 Men;

10. Receipts at Chicago to-day. Wheats 206 cars; corn fill; oats, 181 cars. Estimated receipts to-morrow. Wheat, 400 cars; corn, 85 cars; oats, 235 cars; hogs, 16,000 head.

11. There is a theory evolved by a French scientist to the effect that tho human. race is diminishing In size and will finally become microscopic and vanish into thin air. He says that statistics from the days of the giants to the present time prove that man is getting smaller and shorter and more diminutive live in every way.

12. Red Blankets $1.98 a pair. White Blankets 69c a pair.Bed Comforts 69c each. Heavy Knit Skirts 69c each.

Advice was used at the beginning of the 18th century for military information. The fact that the event takes place in Europe, *Givet*, and what is a small amount of troops for modern times is mentioned, plus the whole linguistics register of the text determines clearly the date of the text in Example (9). As displayed in example (10), the spelling, *to-day* and *to-morrow* is a characteristics of the period between 19th and 20th century, and the quantity involved shows that indeed the time stamp is about that time. The scientific language used in example (11), especially the term *statistics* shows that the text cannot be produced earlier than the second half of the 19th century, yet the mentioning of *days of giants* shows clearly that the science was not yet fully evolved and it was still tributary to an ecclesiastical view of the world. Thus, the text must have been produced around the last quarter of 18th century. The prices specified in example (12) are clearly related to an epoch when the American dollar had a very high value, but yet, it has to be close enough to the modern times in order for an advertisement to the *bed comforts* to be made.

The examples above, which are prototypical for task 2, show that in order to identify correctly the time interval a system must corroborate different types of information, among which an important role play the linguistics register and the details specific to each epoch. In fact, there are few NLP systems, if any, which are able to identify and cluster accordingly to these features. This is why our main effort was directed to provide a good coverage of diachronic corpus especially for task 2, see next section. As we worked on compiling the data for task 2 it becomes clear that a different accuracy is to be expected between task 1 and task 2, and consequently, different types of intervals must be provided for the two subtasks.

The focus of subtask number three is on individual phrases in context. There are certain phrases that are time specific. In fact we can distinguish two categories of such phrases: (i) phrases that have been used preponderantly in a certain epoch and (ii) phrases that have a specific meaning within a certain epoch. For the first type, it is sufficient to recognize them, while for the second, a deeper analysis is necessary and the context in which they are used is relevant. A system able to deal with the challenges posed by task 1 and, especially task 2, must be able to correctly make the distinction between phrases, which carry temporal value and those which do not.

Task 3

13. According to Advices from Germany, a Rupture between the *Courts of Dresden and Berlin* is at Hand

14. The Regiments of Guelderland, and another belonging to this *Republick*, which were accused to not charging the Enemy

15. *corporal punishment*

16. his *artillery* retreat so that he constantly marched under the *grapeshot*

For the contiguous phrases marked with italic format in the examples (13)-(15), a system must be able to decide whether, in the provided context, there is temporal information attached to them. The context is crucial, because, out of context, the temporal value may be cancelled. In a sentence, more than a phrase can be proposed. Roughly, all the features discussed above for task 1 and task 2 are present in the examples of task 3. From this point of view, task 3 can be viewed as a classical feature selection task.

3.2 News Corpus and Data Statistics

Instead of considering whole pieces of news, we focused on individual parts of text that may carry relevant time information. The data proposed for training and test is made out of snippets of text of variable length. Typically a snippet will have between tens to a couple of hundred of words.

We used a series of journals available in electronic format from extracting the data. Most of the electronic archive do not make available newspapers that are older than the beginning of 19th century. However, we wanted to cover the whole period between 1,700 to 2,010. A second detail to consider is the diversity of the sources. Most of the archives are linked to one journal, which restricts the scope of the news to one location and one community. Another aspect that we want to consider for our data is to be hard to find it by searching the web. That will kind of prohibiting a simple system that only does string match to correctly solve the task. A system that find the whole piece of news and its publishing date on the web , may produce good results for task 1 , but would fail to do so for task 2 and task 3.
In order to cope with this restriction we subscribed to several web newspaper archives. The influence of each of these sources in our data set is specified in Table 1.

Source	*address*	*Data task coverage*
NPA	newspaper.achive.com	75%
SPR	archive.spectator.co.uk	12%
BDY	www.bodley.ox.ac.uk/ilej/	10%
OTHER		3%

Table 1: Data Sources.

The separation of the data into trial, training and test is presented in table 2. The data for task1 is not very rich. This is because the learning methodology for task 1 is pretty clear, so we are mainly interested in having a statistical sufficient pool for drawing accurate conclusions after the evaluation of the task. For task 2 the methodology is still a matter of research we want to provide as much data as possible in order for machine learning systems to be able to learn both the surface and meta-textual features. For task 3, there is no need for training. A phrase is or it is not time relevant, and each case must be treated separately.

data	*task 1*	*task 2*	*task 3*
trial	17	87	7
training	167	5, 436	NA
test	267	1, 041	108
total	451	6, 568	115

Table 2: Data size.

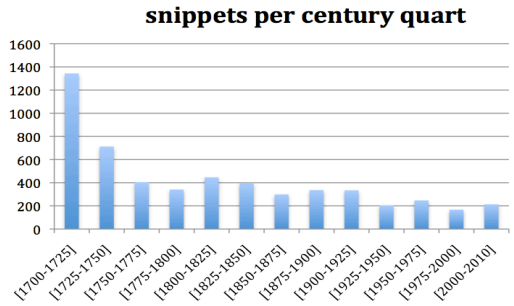

Figure 1: Task 2 distribution.

The snippets cover the last three centuries. However, the number of snippets per year may vary. In Figure 1 we plot the distribution of the number of snippets for each time interval of 25 years for task 2. With the notable difference of the first 50 years of the 18th century, each quarter of the century is covered by a number between 200 to 400 snippets, which men an average between 4 to 8 snippets per year. The first two quarters of the 18th centuries are substantially better covered: 1, 343 and 780 snippets respectively. The explanation for this skewness is two fold: (1) the data for the beginning of the 18th century is much more difficult to acquire than the rest of the data. Basically the text exists only as pdf and the OCRss are not trained to work on this kind of text. Therefore, it is really hard to get a good corpus for the beginning of 18th century, but, as this is in fact our goal, we pursued into acquiring the snippets for this period with priority. (2) the data at the beginning of the 18th century is the one which has a rich variation of linguistic constructions, and the present corpus can be used further for different analyses. We note here, that from the point of view of lexical variability, the 19th century is very rich and there is a huge jump from the previous century in the size of vocabulary.

In this section we have defined the broad scope of the DTE task, we have reviewed the main characteristics of the subtasks, and we have shown to

what type of information must be extracted and managed by a diachronic system. In the next section we present the details of task organization - the format of data, the input and expected output and the evaluation procedure.

4 Task Organization

4.1 Data Format

As this task is the first of its genre, it is hard to know priorly how accurate a system can be in determining epochs and sub epochs from a news corpus. On the basis of our previous experience (Popescu and Strapparava, 2014; Popescu and Strapparava, 2013), we have reasons to believe that separation into epochs is not linear: the epochs tend to change much faster in modern times. However, the topics seemed to be much better differentiate a couple of hundred of years ago than in the modern times. All in all, it seems that a 50 years time interval is something that could be inferred without carrying out a special analysis for both tasks T1 and T2. Thus, in order to be able to judge justly the contribution to each system, a shorter time interval should be taken into account. We have decided to consider an interval centered around the year in which the news was actually produced and to have three types of intervals: fine, medium and coarse. The three intervals are included one in another, and for all three there is an equal number of years to the left and to the right of the actual date. This condition creates intervals with even number of years. We considered the intervals for task T1 and T2 as presented in Table 3.

accuracy	*task1*	*task2*
fine	2	6
medium	6	12
coarse	12	20

Table 3: Time intervals.

The system has to choose the correct time period, e.g. 1700-1720, ..., 1900-1920, ..., from the given set of contiguous intervals which cover the whole period considered, i.e. from 1700 to 2014. In both subtasks 1 and 2 the explicit choice of intervals is available. Only one interval is correct for each level of accuracy. In the training data each snippet has an unique ID, followed by three lines, one for each level of precision and each containing the set of intervals with the specific length. Only one interval is marked with *yes* in training, while in test all are marked with *no*. At the evaluation time, the system performances are compared against the gold standard.

4.2 Evaluation and baseline

The results on each snippet can be evaluated individually. The system has to specify the chosen interval, and if this is the same as the one specified in the gold standard, then the answer is correct, otherwise not. However, the distance from the chosen interval and the correct interval is relevant. Between two systems that have exactly the same number of strictly correct answers, it is preferable to work with the one that has the minimal error average. Keeping in mind these ideas we implemented an evaluation script, which takes into account the distance between the chosen and the gold standard interval. The score is normalized to [0,1) interval. The correct answer is marked with a zero loss and a ten or more interval difference is marked with 0.99 loss. According to the number of intervals off, a loss is computed between 0 and 1, see Table 4. The final score is 1- loss. The evaluation script also outputs the number of years by which the system was off and their distributions, that is, the distribution of loss function from 0 to 9.

We have considered a simple baseline, that is random choice. Another candidate is to always choose the median interval, like 1850, for example. However, both options are bad, and the number of 9 or more intervals off is very large, these baselines tend to have a very high loss function. Their behavior is not actually very different one another. That is why we choose officially to have just one baseline, namely random choice. This choice is supported by the following reason: the median produces every time the same output, while the random choice is different. Averaging over several runs of the random choice we have a much better approximation of what are the baseline performances.

intervals off	*loss*	*intervals off*	*loss*
0	0	5	.5
1	.1	6	.6
2	.15	7	.8
3	.2	8	.9
4	.4	≥ 9	.99

Table 4: Loss as function of off intervals.

5 Submitted Runs and Results

There were 7 teams that expressed their interest in the task, but in the end there were only four teams which successfully submitted the results. The number of submitted runs was less, though, as not all the teams participated in all the tasks. In fact there is only one team that participated in all three tasks, namely **IXA**. As such, we are glad to acknowledge them as the winner of the tasks, if the average over the all three task is made.

We are going to present the team by including their own description of their systems. More details can be found in their system paper, submitted to the SemEval 2015. Then we present their results and discuss the performances of their system individually.

5.1 Systems

A short description of the system follows:

I **AMBRA**

Our approach is based on the learning-to-rank framework using pairwise comparisons, previously proposed for temporal text modelling by (Niculae et al., 2014). 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, 1998). In light of this, our system is named AMBRA (Anachronism Modelling by Ranking). We used four types of features: document length meta-features, stylistic, grammatical, and lexical features. The four stylistic features used were previously proposed by (Stajner and Zampieri, 2013): Average Word Length (AWL), Average Sentence Length (ASL), Lexical Density (LD) and Lexical Richness (LR).

II **IXA**

Four different approaches are undertaken in order to automatically determine the period of time in which a piece of news was written: the first approach consists of searching for the mentioned time period within the text. The second approach, on the other hand, consists of searching for named entities present in the text and then establishing the period of time by linking these to Wikipedia. The third approach uses Google NGrams and, to conclude, the fourth approach consists of using linguistic features that are significant with respect to language change in combination with machine learning.

III **UCD**

We approach the task of dating a text (subtask 2) as a stylistic classification problem. For each level of granularity (6-year, 12-year, and 20-year), we train a multi-class SVM classifier using a set of stylistic features extracted from the texts. These features include frequency counts of character, word, and POS-tag n-grams, and syntactic phrase-structure rule occurrences. We also incorporate date estimates of syntactic nodes from the Google syntactic n-grams database. Our submission is a classifier incorporating all of these features and trained on the task training data. We find that of the stylistic features, character n-grams are the most informative. The Google syntactic n-gram dates, while weak predictors on their own, are also among the most informative features in our combined classifier.

IV **USAAR**

We built a crawler to crawl the text snippets in the task and also we found that the webpages retrieved were dated. We use those dates as answers to the task evaluation. We then crawl the two webpages fully and then clean the website to produce a corpus of diachronic texts for future use (in total 24,280 articles).

5.2 Evaluation

The results are presented in Table 5. The *acc* column lists the score of the system, computed as described in Section 4.2, and the *P* shows how many times the system was perfectly accurate, that is, it found the exact interval. The fine grade seems to be a problem for the big majority of the systems. The only system which reports very high value, USAAR, is

System	*Task 1*						*Task 2*						*Task 3*
	F		M		C		F		M		C		
	acc	*P*	*acc*	*P*	*acc*	*P*	*acc*	*P*	*acc*	*P*	*acc*	*P*	*acc*
AMBRA	.167	.037	.367	.071	.554	.074	.605	.143	.767	.143	.868	.292	NA
IXA	.187	.02	.375	.041	.557	.090	0.261	.037	.428	.067	0.622	.098	.573
UCD	NA	NA	NA	NA	NA	NA	0.759	.463	.846	.472	0.910	0.542	.551
USAAR	.953	.910	.972	.928	.981	.943	NA	NA	NA	NA	NA	NA	NA
baseL	.107	.112	.174	.187	.377	.037	.224	0	.391	0	.524	0	.237

Table 5: DTE results.

based on web crawling, thus is not a generalizable method. In fact, the team participated only in task 1. The medium grade seems to be doable, all systems scoring better than the baseline. For the coarse grade the systems outperform the baseline by several tens of percent and obtain very good results, with accuracy between 0.868-0.91. These results confirm the fact that the task is doable and a 20 years interval is appropriate for DTE. We hope these results can be further improved in the future.

The results for task 3 show that this task is indeed difficult, and even if the baseline has been overcome with a great margin, the results show that the system could be improved further. We plot the distribution of errors for the system which participated in task 2, see Figure 2. Interestingly, AMBRA and UCD have very similar distributional curves, with the exception of perfect guess. The IXA system has a more regular shape and its errors seem to be evenly distributed with a big exception for the maximum error category. Maybe an interpolation between these three methods could lead to a better overall result.

To conclude, we are glad we received different systems which produce good and very good results. These initial ideas represent a valuable pool from which further work can be developed in the future.

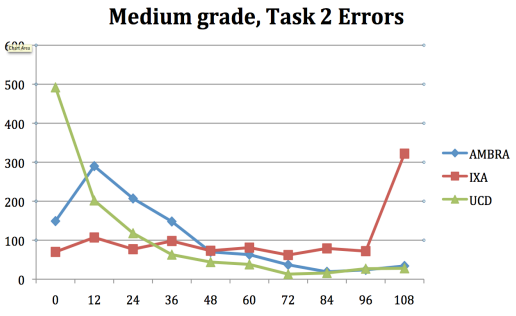

Figure 2: Task 2 medium error distribution.

6 Conclusion and Further Research

In this paper we described the Diachronic Text Evaluation task. We explain the main motivation for this task and we presented what the main issues behind the diachronic task are and how these issues have influenced our decisions. We presented the sources and the distribution of snippets in task data. A short paragraph description for each of the participating systems is provided, and we carried out a global evaluation. Finally we have provided an analysis of errors for task 2.

We think that there are some very interesting directions we would like to investigate further. The first one is to consolidate the actual corpus. This a necessary step in order to build a solid basis for further experiments and developments. We would like to improve the quality and quantity of training text for allowing search of changes at all linguistics level. We would like to work more in revealing the connection between diachronic evaluation and epoch discovery.

Another direction of research is a systematic study of the textual and meta-textual features that are relevant for the DTE task and what their individual contributions to the over all accuracy is. Besides the overt temporal features we need to identify, the linguistics register, the topics and the discourse features - from grammar to pragmatics must be taken into account. We believe that DTE is a very good indicator on the performance of machine learning systems for the meta-textual feature management.

Last, but not least, we would like to bridge the gap between different old and emergent fields, such as sociology, socio-historic linguistics and social computational analysis, computational journalism and forensic linguistics respectively. We think that NLP systems are able to tackle the difficult issues posed by this research.

References

Thorsten Joachims. 1998. Text categorization with Support Vector Machines: learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137–142.

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

Octavian Popescu and Carlo Strapparava. 2013. *Behind the Times*: Detecting epoch changes using large corpora. In *Proceedings of the 6^{th} International Joint Conference on Natural Language Processing (IJCNLP-2013)*, Nagoya, Japan, October.

Octavian Popescu and Carlo Strapparava. 2014. Time corpora: Epochs, opinions and changes. *Knowledge-Based Systems*, 69:3—13, October.

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)*.