IXAGroupEHUDiac: A Multiple Approach System towards the Diachronic Evaluation of Texts

Haritz Salaberri†, Iker Salaberri‡, Olatz Arregi†, Beñat Zapirain†
†*IXA* Group - Faculty of Computer Sciences, ‡Faculty of Arts
University of the Basque Country, Spain
firstname.lastname@ehu.eus

Abstract

This paper presents our contribution to the SemEval-2015 Task 7. The task was subdivided into three subtasks that consisted of automatically identifying the time period when a piece of news was written (1,2) as well as automatically determining whether a specific phrase in a sentence is relevant or not for a given period of time (3). Our system tackles the resolution of all three subtasks. With this purpose in mind multiple approaches are undertaken that use resources such as Wikipedia or Google NGrams. Final results are obtained by combining the output from all approaches. The texts used for the task are written in English and range from the years 1700 to 2000.

1 Introduction

According to Mihalcea and Nastase (2012) current applications within human language technology work with languages as if they were constant. However, changes in language are taking place constantly, for example: new meanings for old words are coined; metaphoric and metonymic uses become so ingrained that they are considered literal from one specific point in time on; new words are constantly being created.

These changes in language are what in part has motivated the task addressed by our system. In fact, subtasks (1) and (2) tackle the problem of computationally identifying the time period in which a piece of news was written. This is undertaken based on, among other things, the changes that take place in language over time. The difference between subtasks (1) and (2) is that the texts in subtask (1) contain clear references to time anchors. This means that e.g. historical events, relevant people, commercial products etc. are mentioned in the text that are specific to the period of time in which the texts were written. Subtask (3), on the other hand, consists of determining whether a phrase within a clause is specific or not to the period of time in which the text was written. The training corpus for this subtask is made up of the texts from other subtasks. As a consequence our system will be able to use information on both time anchors and language changes in order to generate the results for subtask (3).

This paper is organized as follows: section 2 presents the resources available to the diachronic evaluation of texts; section 3, on the other hand, shortly reviews the relevant literature on this matter. Section 4 makes a description of the developed system; results are then described in section 5 and, finally, our conclusions are given in section 6.

2 Resources

To the extent of our knowledge, there exist two main resources as of today for computationally addressing the diachronic evaluation of texts as defined in task 7: Google NGrams and Wikipedia. The former holds statistics on word usage on Google Books, a textual corpus consisting of books written in English and printed between 1505 and 2008. Google NGrams can be used to map language changes to specific time periods. The latter requires no presentation as it is a well-known resource; it can be used to establish the period a time anchor belongs to.

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 840–845,
Denver, Colorado, June 4-5, 2015.

3 Related Work

To the best of our knowledge several techniques have been previously used to computationally address language-change. We consider it important to note that the motivation to study the language-change phenomena differs from one work to another: Some of the techniques make use of it in order to establish the period of time in which a text was produced (Jong et al., 2005; Dalli and Wilks, 2006), which is our main concern; others, on the other hand, use the phenomena in order to study topics such as the changes that have taken place in culture (Juola, 2013; Michel et al., 2011).

Some of the techniques used so far to address the task of temporal classification are based on language models built from texts belonging to a same period of time. This way the task of temporally classifying texts consists basically of identifying the model that best fits the text that wants to be classified. Some of the systems that follow this approach are Kumar (2011) and Wang et al. (2012).

Another relevant class of models for temporal classification is based on the idea that the change of word meaning and word usage over time can help determine the period of time in which a text was written. Normally the resource used by the systems that are based on this approach is Google NGrams (see section 2). Some example models that use this approach are presented in Mihalcea and Nastase (2012) and Popescu and Strapparava (2013).

Other systems that can be brought up in this section make use of stylistic and readability features (Štajner and Zampieri, 2013), neural nets (Kim et al., 2014) and lexical features (Dalli and Wilks, 2006).

From the approaches here presented we decided to implement our system using, among others, the change of word usage and word meaning over time approach (see subsection 4.1.3) and the lexical and stylistic features approach (see subsection 4.1.4) as we believe both to have reported good performance in previous works (Mihalcea and Nastase, 2012; Štajner and Zampieri, 2013; Dalli and Wilks, 2006). Although we think that the approach to epoch delimitation based on using language models can also come up with good results, we have not used it as we believe that the training set is too limited for this approach to be effective.

4 System Description

The way in which our system deals with temporal text classification (subtasks (1) and (2)) is described under subsection 4.1. The way in which our system deals with recognizing time-specific phrases (subtask (3)), on the other hand, is presented under subsection 4.2.

4.1 Temporal Text Classification

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.

4.1.1 Year Entity Detection

The present approach was implemented based on the observation made upon the training texts, in the development of which we have realized that the period of time that corresponds to a text is present within the text. This approach is characterized by a very high precision and a very low recall as only 10% of the training texts contain a period of time and in 85% of the cases these are the ones that correspond to texts. In order to establish the time period, year entities are detected by our system using the Apache OpenNLP name finder tool Baldridge (2005).

It is considered here that this approach is strongly dependent on the domain; in fact, if historical texts (or texts that in general describe past events) were to be diachronically evaluated, the precision would drop and recall would improve considerably.

4.1.2 Wikipedia Entity Linking

For the second approach our system detects named entities that correspond to persons and organizations within the texts; the Apache OpenNLP name finder tool and the pre-trained models for this

type of entities are used. After named entities are recognized, these are searched for in Wikipedia; if a named entity can be found, then year entities are detected in the corresponding entry: with this purpose in mind the OpenNLP name finder tool is used and tuned as in 4.1.1. Finally, an average of all years (which stem from the Wikipedia entries that correspond to the named entities in the text) is calculated for every text and the time period that corresponds to the average assigned. The workflow for this approach can be seen in figure 1:

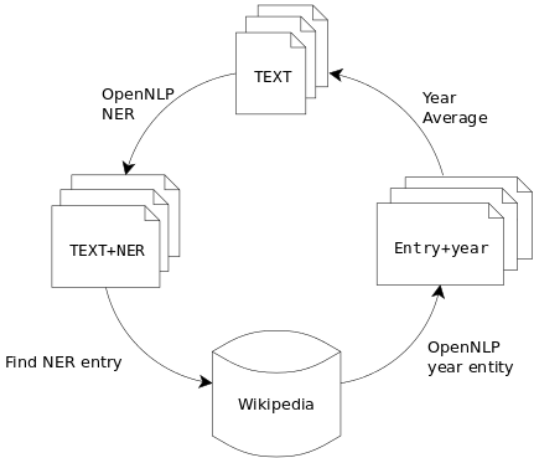

Figure 1: Wikipedia Entity Linking.

Different scenarios are possible concerning this approach: some texts do not contain named entities and some others have many of them, and sometimes entities are not detected or can not be found in Wikipedia. For these reasons not all texts are assigned a time period by this approach[1].

4.1.3 Google NGrams

The Google NGrams 1-gram corpus is used for the third approach. We consider all nouns (proper and common) within the texts to be of interest as we consider these to be the kind of words that change most across time and as a result provide the highest amount of information on the time in which a piece of news was written. In order to identify these nouns the ClearNLP PoS tagger and lemmatizer is used Choi and Palmer (2012). The system computes for each noun the percentage of occurrences that that noun has in a year with respect to the sum of words available for that year (normalization). The amount of published data in Google Books is not the same for all years; in fact, it grows exponentially from the second half of the 20th century on. For this reason the percentage of occurrences with respect to the sum of words needs to be calculated, rather than simply using the amount of occurrences.

When percentages for all nouns in a text are calculated, the year that corresponds to the highest percentage is associated to each noun. Then, the average value for these years is calculated. If the year associated to a noun differs in 40 or more years from the average value, our system considers this noun to be period-specific. Consequently, the time period that includes this year is assigned to the text. Period-specific nouns are determined locally within a given text since the same noun might be period-specific in one text but not in another.

If there is more than one noun that is considered to be period-specific, the average value of the years that correspond to these nouns is used. If there are no detected period-specific nouns, on the other hand, the average value calculated for all nouns is used.

4.1.4 Language Change

The fourth approach used by our system consists of using linguistic features (patterns or tendencies) that are significant regarding language change in combination with machine learning. For this purpose the different diachronic or language-change tendencies that are observable in the training data have been studied. These tendencies include both linguistic and extra-linguistic factors, and they affect different areas of grammar such as orthography, lexicon, semantics, morphology or syntax. Some examples can be seen in figure 2.

The patterns resulting from the study are classified into six different fifty-year periods ranging from the years 1700 to 2000 as we consider these to be the finest grain period patterns can be classified into. Said patterns are used as features for the learning algorithm; some examples include: the loss of subjunctive mood in subordinate clauses, the arisal of *do so*-verbal substitution and the extinction of postpositions and of various inflectional morphemes. In spite of the richness of extracted linguistic change patterns, this approach has proved in any case to be

[1]Our approach does not handle cases where more than one Wikipedia pages match a name.

ORTHOGRAPHY

- Until about 1720 reflexive pronouns and their possessors are written separately instead of together:

...she fancies her self in a Wood... (1707-1713)

- About 1895 the contraction Messr. is replaced by Mr. for 'mister' or 'messieur':

...from Messrs. Chatto & Windus... (1894-1900)

...Mr. Balfour appears in the strange capacity... (1904-1910)

LEXICON

- Use of the archaic locative adverb 'thither' in the sense of 'in this direction, here'. 1720 at the latest:

...the reinforcements sent thither from Milan and Spain... (1698-1704)

- Loss of the word guineas for pounds, around 1900:

...lowest price 130 guineas... (1814-1820)

...just pays a couple of pounds... (1970-1976)

SEMANTICS

- Use of the verb 'fit' in the sense of 'walk, move' instead of 'to suit', until around 1715:

...as I fate under the Shadow of it... (1706-1712)

- The verb 'to wit' is used in the sense of 'to know' in a phossilized expression. These verb and expression are no longer used nowadays.

...That afterwards, to wit, on the twenty seventh day... (1715-1717)

MORPHOLOGY

- Fossilized trace of 2nd person singular marker in verbal morphology: (you) Could'st instead of you could, 1710 at the latest.

...Jerusalem! Could'st thou but know... (1699-1705)

- Cliticization of the pronoun 'it' to the conjugated copula 'is' into 'tis'. Until about 1720, although frozen uses could exist even today.

...Tis such an Entertaiment... (1707-1713)

SYNTAX

- Complete loss of the non-do-auxiliary pattern of the do-auxiliary in negative clauses, emphatic constructions, and yes/no questions, approximate date 1730. This implies loss of the pattern of the negative particle following the finite verb..

...and I hear of I know not how... (1709-1715)

Figure 2: Some of the language-change patterns used by our system.

much less effective when compared to the other approaches.

The classifier used by the approach here described is a standard multi-class *Support Vector Machine* classifier implemented using the *SVM-multiclass* package in Joachims (1999). The decision of using a standard SVM learning algorithm comes from our experience on classification tasks with such a large number of classes.

4.1.5 Final Decision

In order to ultimately determine the period of time in which a text was written the system follows a procedure that takes into account the precision given by each approach (since the systems seeks maximum precision). We consider the year entity detection approach to be the one with the highest precision, followed by the Wikipedia entity linking, the Google NGrams and the language-change approaches. The present procedure establishes that the period of time yielded by the approach with the maximum precision that is available must be set to the text. It must be kept in mind that both the year entity detection and the Wikipedia entity linking approaches have a low recall as only some of the texts are assigned a period of time by these approaches.

Subtask \ Grain	Coarse		Medium		Fine	
	Precision	Score	Precision	Score	Precision	Score
1	0.0902	0.5575	0.0413	0.3672	0.0225	0.187
2	0.0987	0.6225	0.0677	0.428	0.0377	0.2618
3	0.5739					

Table 1: Official results reported for our system for all three subtasks.

4.2 Recognizing Time-Specific Phrases

We consider that determining whether the phrases within a sentence are particularly relevant or not for the period of time in which the sentence was written can be viewed as a two-step procedure: first, markable phrases need to be detected, and then it must be decided whether these phrases are indicative features for the period of time or not. Our system performs just the classification step since the markable phrases are provided by the task organizers. This is achieved by making use of the period-specific words identified in the Google NGrams approach described in 4.1.3. Our system marks the set of consecutive words that start and end with period-specific words as a relevant phrase for the period of time in which the text was written. This procedure is followed if there is no punctuation mark between the words and the distance is not greater than four words.

The decision to consider phrases that have a maximum of four words is based upon observation. We consider this to be the appropriate number of words in order not to miss too many relevant phrases. The system can be easily tuned for phrases with a greater or a smaller number of words.

5 Results

Table 1 contains the official results reported for our system. In order to evaluate subtasks (1) and (2) three configurations are considered: a fine-graded evaluation were periods of time span two years in subtask (1) and six years in subtask (2); a medium-graded evaluation were periods of time span six years for subtask (1) and twelve years for subtask (2) and a coarse-graded evaluation were periods of time span twelve years in subtask (1) and twenty years in subtask (2).

There is no fine-, medium- or coarse-graded evaluation for subtask (3). Certain phrases from a piece of news are selected by the task organizers and marked as *yes* or *no* by our system according to their relevance for the period of time when the news was produced (the period of time is also provided by the organizers). The score for this subtask is computed by counting the number of times our system has correctly marked the phrases.

As far as we know the only works that bear a slight resemblance to what is proposed in the temporal text classification subtasks (subtasks (1) and (2)) are Mihalcea and Nastase (2012) and Popescu and Strapparava (2013), in which computational approaches to temporal classification of words are presented. We consider that our results cannot be even loosely compared to the results in the cited papers as there is too little resemblance between temporal text classification and temporal word classification. We are not aware of any work that performs recognition of time-specific phrases (subtasks (3)).

As can be observed in table 1, the scores for subtask (2) are higher than the scores reported for subtask (1); however, we find that establishing the period of time when a piece of news was written is more complicated for the texts in subtask (1) as it mainly depends on a correct exploitation of time anchors. For this reason, we understand that the performance of our system is higher in subtask (1) than in subtask (2). Finally, we believe that the score obtained for the third subtask (0.5739) can be understood as an indicator of high performance as the difficulty of the subtask is, in our opinion, higher than that of other subtasks.

6 Conclusions and Future Works

In this paper we have presented our system for the diachronic evaluation of English texts, which has taken part in the SemEval-2015 task 7. Our system has been the only participant system that has reported results for the three subtasks that comprehended the task. We believe that many issues still

need to be reviewed.

We intend to improve the overall performance of the system in the near future by trying out new techniques that we have not been able to implement due to time limitations.

Acknowledgments

Haritz Salaberri holds a PhD grant from the University of the Basque Country (UPV/EHU)(*IXA* Group, Research Group of type A (2010-2015)(IT34410)). In addition, this work has been supported by the FP7 *NewsReader* project (Grant No. 316404).

References

Jason Baldridge. 2005. The opennlp project. *URL: http://opennlp. apache. org/index. html,(accessed 2 February 2012)*.

Jinho D. Choi and Martha Palmer. 2012. Fast and robust part-of-speech tagging using dynamic model selection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 363–367. Association for Computational Linguistics.

Angelo Dalli and Yorick Wilks. 2006. Automatic dating of documents and temporal text classification. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 17–22. Association for Computational Linguistics.

Thorsten Joachims. 1999. Making large scale SVM learning practical. *Universität Dortmund*.

F. de Jong, Henning Rode, Djoerd Hiemstra. 2005. Temporal language models for the disclosure of historical text. *Royal Netherlands Academy of Arts and Sciences*.

Patrick Juola. 2013. Using the Google N-Gram corpus to measure cultural complexity. In *Literary and linguistic computing 28(4)*, pages 668–675. Association for Computational Linguistics.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, Slav Petrov. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

Abhimanu Kumar, Matthew Lease, Jason Baldridge. 2011. Supervised language modeling for temporal resolution of texts. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2069–2072. Association for Computational Linguistics.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant and others. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 259–263. Association for Computational Linguistics.

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

Octavian Popescu and Carlo Strapparava. 2014. Time corpora: Epochs, opinions and changes. *Knowledge-Based Systems*.

Sanja Štajner and Marcos Zampieri. 2013. Stylistic changes for temporal text classification. *Text, Speech, and Dialogue*, pages 519–526. Springer.

Chong Wang, David Blei, David Heckerman. 2012. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*.