USAAR-CHRONOS: Crawling the Web for Temporal Annotations

Liling Tan and Noam Ordan
Universität des Saarlandes
Campus A2.2, Saarbrücken, Germany
alvations@gmail.com, noam.ordan@uni-saarland.de

Abstract

This paper describes the USAAR-CHRONOS participation in the Diachronic Text Evaluation task of SemEval-2015 to identify the time period of historical text snippets. We adapt a web crawler to retrieve the original source of the text snippets and determine the publication year of the retrieved texts from their URLs. We report a precision score of >90% in identifying the text epoch. Additionally, by crawling and cleaning the website that hosts the source of the text snippets, we present Daikon, a corpus that can be used for future work on epoch identification from a diachronic perspective.

1 Introduction

"*Time changes all things: there is no reason why language should escape this universal law*" (De Saussure, 1959). Traditionally, there are two ways to collect linguistic data to explore how words change over time, viz. (i) the 'armchair' method and (ii) the 'tape-recorder' method (Aitchison, 2001). In the first, the linguist cross-examines numerous documents from bygone years and in the latter, the linguist goes around recording language and studies the changes as they happen.

With the ingress of historical data provided by Google (Michel et al. 2011), the 'armchair' method goes into warp speed as computational linguists explore the different facets of lexical changes in English (Mihalcea and Nastase, 2012; Popescu and Strapparava, 2013; Niculae et al., 2014).

This paper presents the Saarland University (USAAR-CHRONOS) participation in the Diachronic Text Evaluation task in SemEval-2015. We participated in Subtask 1 that requires participants to identify the year of publication for texts with clear reference to time anchors (i.e. explicit references to famous persons or events).

1.1 Task Definition

In Subtask 1 of the Diachronic Text Evaluation participants are required to identify the epoch (i.e. time period) of a text snippet with clear reference to certain famous persons or events. The text snippets may not necessarily contain temporal information such as year or date but it has clear reference to a historical event that can be identified from external knowledge bases. For instance, given the following text, participants are required to identify its epoch:

"*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.*"

The text has clear temporal evidence with reference to a historical figure ("*Saddam Hussein*"), a notable organization ("*UN Security Council*") and a factual event ("*Operation Desert Shield*"). Historically, we know that Saddam Hussein lived between 1937 to 2006, that the UN Security Council has existed since 1946 and that Operation Desert Shield (i.e. the Gulf War) occurred between 1990-1991. Given the specific chronic deicticity ("*today*") that indicates that the text is published during the Gulf

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 846–850,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

War, we can conceive that the text snippet should be dated 1990-1991.

For each text snippet, different epoch choices are provided at three granularity levels; fine, medium and coarse graded epochs, and they are assigned the time periods of 3, 6 and 12 years, respectively. For the given example above, the correct epochs are 1990-1992, 1988-1993 and 1985-1995 for the three granularity levels respectively.

2 Related Work

Michel et al. (2011) launch the field of *culturonomics* to study changes in human culture through language change; for this, they release ngrams taken from millions of digitized books; they show, for example, that censorship and suppression can be determined by comparing the frequencies of proper names in multilingual ngrams in this dataset.

Mihalcea and Nastase (2012) explore word sense disambiguation over time using snippets from Google Books; they add a semantic dimension on top of lexical frequency to conduct word epoch disambiguation based on the fact that words change their neighbors throughout time.

The Google Ngram corpus has spawned several related studies. To create a sense pool, Yu et al. (2007) extract pairs of ngrams and filter them with an appropriate statistical test using their frequencies, where the resulting sense pool is manually verified. Interestingly, their experiments conflate the ngrams across time, yet it is unclear whether the resulting sense pool contains ngrams across different epochs. Juola (2013) uses the bigrams from the Google Books Ngram dataset to measure changes in the Kolmogorov complexity of American culture at ten-year intervals between 1900 and 2000. Related to this, Štajner and Zampieri (2013) show, for Portuguese, that lexical richness, average word length and lexical density increase over a span of 400 years.

Topic models are also applied to study topical changes across epochs (e.g. (Blei and Lafferty, 2007; Wijaya and Yeniterzi, 2011)). Related to epoch identification, Wang and McCallum (2006) develop time-specific topic models to a time stamp prediction task.

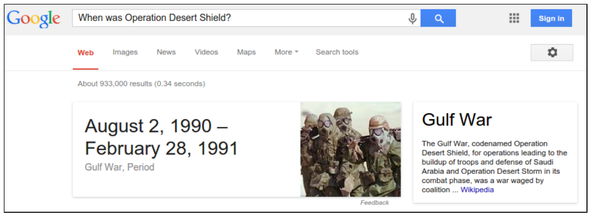
With the renaissance of neural nets, recent studies are using deep neural language models to detect diachronic lexical changes from several text types ranging from published books (Kim et al., 2014) to Twitter microblogs and Amazon movie reviews (Kulkarni et al., 2014).

3 Approach

We take a different approach compared to previous studies that treat epoch identification as a classification task. We see it as an information retrieval task where we want to know whether we can get the temporal information of the text snippets from the Internet.

In the age where there is a contest (known as "Googlewhack") for finding one-hit results on Google since they are so rare , it is clear that a great deal of the information we are looking for is just "out there" for us to search. It is recommended to use machine learning classifiers for cases where test data is supposedly unknown, but more often than not it can be known by those who know how to retrieve, clean and harvest systematically.

Prior to the days of Google and search engines, historians and librarians[1] had to cross-reference history books and newspaper archives to identify the text epoch. The Internet is vast and infinite. Given the advent of Wikipedia and Google, epoch identification can be as simple as searching "*When was Operation Desert Shield?*" on Google[2] (see Figure 1).



When was Operation Desert Shield?

About 933,000 results (0.34 seconds)

August 2, 1990 – February 28, 1991

Gulf War, Period

Gulf War

The Gulf War, codenamed Operation Desert Shield, for operations leading to the buildup of troops and defense of Saudi Arabia and Operation Desert Storm in its combat phase, was a war waged by coalition ... Wikipedia

Figure 1: Google Result for "*When was Operation Desert Shield?*".

Tan et al. (2014a) develop a Web Translation Memory (`WebTM`) crawler capable of harvesting parallel texts from the web given an initial seed corpus, similar to the BootCaT system (Baroni and Bernardini, 2004). They adapt `WebTM` such that it attempts to find occurrences of the text snippets from the web. This is akin to developing a dedicated

[1] With the exception of the polymath librarian, Flynn Carsen

[2] See http://goo.gl/VD2Xtx

search- and crawl-system for the purpose of knowledge extraction.

Surprisingly, the source of the all the text snippets of Subtask 1 is found on http://freepages.genealogy.rootsweb.ancestry.com/∼dutillieul and http://archive.spectator.co.uk/. Moreover, these webpages contain dates in their URL, so we extract the publication year with regex pattern matching. Since the task requires an epoch (time period) instead of a discrete publication year, we perform some minor integer manipulation to fit the publication year to the expected epoch[3].

4 Results

Out of the 267 text snippets, our system correctly identifies 243, 248, 252 epochs for the fine, medium and coarse epoch granularities.

	Fine	Medium	Coarse
AMBRA	0.0374	0.0711	0.0749
IXA-EHUDIAC	0.0225	0.0413	0.0902
USAAR-CHRONOS	**0.9288**	**0.9101**	**0.9438**

Table 1: Precision scores on Subtask 1.

Table 1 presents the precision scores of the participating teams in subtask 1. Our system scores best on all three granularity levels.

Figure 2 shows a heatmap of the fine graded epochal (6 years interval) differences between the outputs and the gold standard[4]. The warm colors indicate higher values within the interval. Looking at the orange region of the heatmap, the other systems were way off in the epoch identification where respectively, AMBRA and IXA-EHUDIAC have 195 and 186 predictions that are 54 years off from the gold standards. We have a total of 24 predictions different from the gold standard and 9 out 24 were 6 years off from the gold standards.

5 Discussion

We have manually checked our epoch predictions and the years encoded in the URL to check whether they correspond to the date of the source articles.

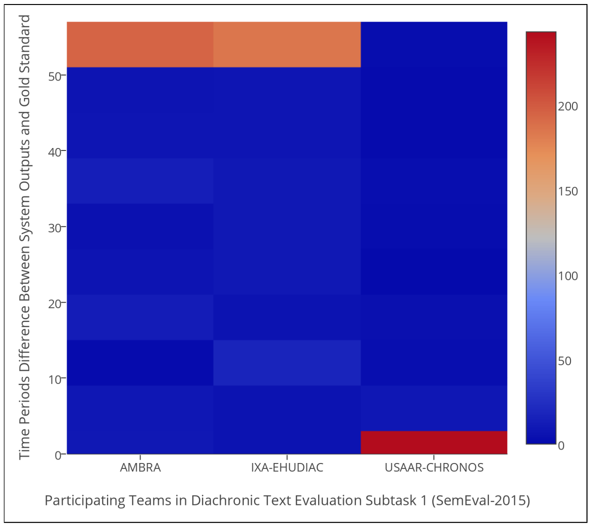


Figure 2: Fine Graded Epoch Differential between Systems outputs and Gold Standards (warmer colors indicates higher values).

Some of our predictions are dated older than the gold standards and vice versa.

For instance, the following text refers to the *Battle of Salamanca* on 22 July 1812 and the text snippet is from a battle report written on 16 August 1812 and published on 24 August 1812 in the *Salisbury and Winchester Journal*; the gold annotation records the epoch as 1813-1815 whereas our system reports 1810-1812.

"*On Thursday last, the 69th Annual Conference of the people called Methodists, was concluded. It had been held by adjournment in Leeds from the 27th ult. About 309 Itinerant Preachers were present from various parts of the United Kingdom, who gave very gratifying accounts of the success with which their ministry have been crowned.*"

In this case, the gold standard source is clearly a different source and the assumption that there are hard boundaries in epoch identification should be relaxed. One should consider different granularity levels of the epochs involved when evaluating the system's accuracy.

Relating to the historian and librarian anecdote, the discrepancy in dates from different sources shows that cross-referencing temporal annotations from various sources should be considered in future diachronic studies and temporal analyses.

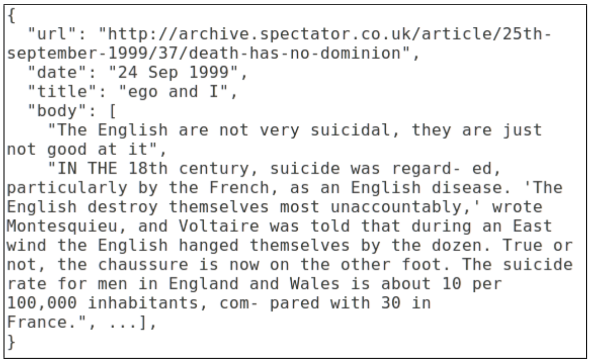
[3] Details on http://goo.gl/TcZ9z0

[4] An interactive version of the heatmap can be viewed on https://plot.ly/ alvations/21/epochs-differential/

6 Daikon Corpus

After the SemEval task, we crawled the full articles from `http://archive.spectator.co.uk/`, cleaned the corpus and annotated it with the exact publication date of the article, its title and the URL from which it was retrieved. The Daikon Corpus is made up of articles from the British Spectator news magazine from year 828 to 2008.

The Daikon corpus can be used for future diachronic studies and epoch identification tasks; it provides a complementary dataset to the gold standard provided by task. The corpus is saved in JSON format. An excerpt from the corpus looks like this:



```
{
  "url": "http://archive.spectator.co.uk/article/25th-
september-1999/37/death-has-no-dominion",
  "date": "24 Sep 1999",
  "title": "ego and I",
  "body": [
    "The English are not very suicidal, they are just
not good at it",
    "IN THE 18th century, suicide was regard- ed,
particularly by the French, as an English disease. 'The
English destroy themselves most unaccountably,' wrote
Montesquieu, and Voltaire was told that during an East
wind the English hanged themselves by the dozen. True or
not, the chaussure is now on the other foot. The suicide
rate for men in England and Wales is about 10 per
100,000 inhabitants, com- pared with 30 in
France.", ...],
}
```

Figure 3: An Excerpt from the Daikon Corpus.

Each item in the body list is a paragraph embedded within the `<p>...</p>` tags of the webpage. The corpus contains 24,280 articles with 19 million tokens; the token count is calculated by summing the number of whitespaces plus 1 for each paragraph.

To clean the corpus, the encodings are converted to Unicode (UTF8) and XML escape tokens are converted to its Unicode counterparts automatically[5]. However, the current version still contains minor tokenization errors such as the hyphenation error seen in Figure 3. Probably, a character language model could be developed to identify lexical items bounded by the r‘`\w+- \w+`’ regex.

7 Conclusion

In this paper, we have described our submission to the Diachronic Text Evaluation for SemEval-2015. We have adapted a web crawler to search for the source of the text snippets used for the evaluation and achieved the highest precision score. Additionally, we have crawled and cleaned the source articles of the snippets and produced the `Daikon` corpus that can be used for future research in diachronic/temporal analysis and epoch identification.

Acknowledgements

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n ° 317471. In addition, we received support from Deutsche Forschungsgemeinschaft (DFG) through grants from the Cluster of Excellence – Multimodal Computing and Interaction (EXC-MMCI) and SFB 1102. We thank Elke Teich for her encouraging input.

References

Jean Aitchison. 2001. *Language Change: Progress or Decay?* Cambridge University Press.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04).*

David M Blei and John D Lafferty. 2007. A Correlated Topic Model of Science. *The Annals of Applied Statistics*, pages 17–35.

Ferdinand De Saussure. 1959. *Course in General Linguistics.* New York:McGrawHill.

Guy Emerson, Liling Tan, Susanne Fertmann, Alexis Palmer, and Michaela Regneri. 2014. SeedLing: Building and Using a Seed corpus for the Human Language Project. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 77–85, Baltimore, Maryland, USA, June.

Patrick Juola. 2013. Using the Google N-Gram corpus to Measure Cultural Complexity. *Literary and linguistic computing*, 28(4):668–675.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. *arXiv preprint arXiv:1405.3515.*

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2014. Statistically Significant Detection of Linguistic Change. *CoRR*, abs/1411.3315.

[5] The cleaning tool used is a compilation of web cleaning scripts (Emerson et al., 2014; Tan et al., 2014b; Tan and Bond, 2011)

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, et al. 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.

Vlad Niculae, Marcos Zampieri, Liviu P. Dinu, and Alina Maria Ciobanu. 2014. Temporal Text Ranking and Automatic Dating of Texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*.

Octavian Popescu and Carlo Strapparava. 2013. Behind the Times: Detecting Epoch Changes using Large Corpora. In *Proceedings of 6th International Joint Conference on Natural Language Processing (IJCNLP)*.

Sanja Štajner and Marcos Zampieri. 2013. Stylistic Changes for Temporal Text Classification. In *Proceedings of the 16th International Conference on Text Speech and Dialogue (TSD2013), Lecture Notes in Artificial Intelligence (LNAI)*, pages 519–526, Pilsen, Czech Republic. Springer.

Liling Tan and Francis Bond. 2011. Building and Annotating the Linguistically Diverse NTU-MC (NTU-Multilingual Corpus). In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 362–371, Singapore.

Liling Tan, Anne Schumann, Jose Martinez, and Francis Bond. 2014a. Sensible: L2 Translation Assistance by Emulating the Manual Post-Editing Process. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 541–545, Dublin, Ireland.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014b. Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection. In *Proceedings of The 7th Workshop on Building and Using Comparable Corpora (BUCC)*.

Xuerui Wang and Andrew McCallum. 2006. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding Semantic Change of Words over Centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, pages 35–40.

Liang-Chih Yu, Chung-Hsien Wu, Andrew Philpot, and EH Hovy. 2007. OntoNotes: Sense Pool Verification using Google N-gram and Statistical Tests. In *Proceedings of the OntoLex Workshop at the 6th International Semantic Web Conference (ISWC 2007)*.