The Power of Temporal Features for Classifying News Articles

Lukas Lange
Saarland University
Saarbrücken, Germany
llange@lsv.uni-saarland.de

Omar Alonso
Microsoft
Mountain View, CA, USA
omalonso@microsoft.com

Jannik Strötgen
Bosch Center for Artificial Intelligence
Renningen, Germany
jannik.stroetgen@de.bosch.com

ABSTRACT

Temporal information extracted from texts and normalized to some standard format has been exploited in a variety of tasks such as information retrieval and question answering. Classifying documents into categories using temporal features has not yet been tried. Such a method might be particularly valuable when classifying sensitive texts such as patient records, i.e., whenever the pure content of the documents should not be used for the classification. In this paper, we describe, as a proof-of-concept, our work on classifying news articles exploiting only features defined over extracted and normalized temporal expressions. Our evaluation of two classification models on large German and English news archives shows promising results and demonstrates the discriminative power of temporal features for topically classifying text documents.

KEYWORDS

temporal information; news classification; temporal n-grams

ACM Reference Format:
Lukas Lange, Omar Alonso, and Jannik Strötgen. 2019. The Power of Temporal Features for Classifying News Articles. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3315000

1 INTRODUCTION

Motivation. Classifying news articles according to a set of categories depending on the articles' content is a standard task. The terms in the texts can be used as weighted features to precisely classify a document as belonging to a particular category such as *politics* or *sports*, e.g., using SVMs [6]. More sophisticated methods than bag-of-words models can be applied, e.g., by using n-grams.

However, in some scenarios, classification needs to be performed without looking deeply into the texts' content, e.g., due to privacy issues when classifying emails or patient records. One approach to avoid using content terms is to apply character-level methods, e.g., character-level convolutional networks [10]. An alternative is to use previously extracted and normalized pieces of information. A particular type of such information is temporal information.

Related Work on Exploiting Temporal Information. With the availability of temporal taggers to extract and normalize temporal expressions from texts [9], more and more applications are exploiting temporal information extracted from text documents. There is little work on using temporal expressions for classifying news articles like our proposed technique. Rocha et al. [7] investigate techniques for context selection and provide an algorithm that identifies which temporal context is best for a classification task. Chambers proposes learning models that use temporal expressions for labeling documents with a correct timestamp [4].

In [3], temporal expressions in documents of a diachronic news archive have been used to analyze how the past is remembered in different countries. In temporal IR, temporal information can be considered as a query topic to specify temporal constraints [1]. The probably most related work to our approach is [2], where search results have been clustered on a timeline, i.e., temporal information was used for classifying documents, however, along timelines and not according to topical categories as in our work.

Contributions. We present our approach to exploit temporal information extracted from texts and normalized to a standard format for topically classifying news articles. Besides the curiosity to determine the discriminative power of temporal features for text classification, the goal of our work is to validate how well texts can be classified only using temporal information extracted from respective documents as an alternative to other methods.

2 METHODS

We use solely temporal expressions and their characteristics to define features for our classification approach.

Temporal Expressions. TimeML (www.timeml.org) defines the `TIMEX3` tag to annotate date (e.g., today), time (e.g., 9 pm), duration (e.g., two days), and set (e.g., weekly) expressions. Its *value* attribute contains normalized information, e.g., about the length of a duration or the anchoring of a date on a timeline. Dates and times can be of different granularities and are realized explicitly (May 2018), implicitly (Black Friday 2018), relatively (next month) or underspecified (April) [9] and refer to the past, present, or future.

Temporal Classification Features. In the following, we describe all features we use for classifying news articles. All features are calculated per document. As basic features (*Types* in Sec. 3), we use

- the numbers of `TIMEX3` annotations per `TIMEX3` type, i.e., number of date, time, duration, and set expressions

For more advanced features, we exploit the *value* attribute of `TIMEX3` annotations. However, fully normalized *value* information, e.g., `2016-11-23` for "November 23, 2016", is very sparse across a document collection. Thus, we generalize the normalized information, e.g., `2016-11-23` as Day and `2018-04` as Month. Based on these generalized *values*, we define the *Value* and Bigrams features:

- a counter for each granularity (e.g., Day, Week, Month, ...)
- temporal bigrams (2 consecutive temporal expressions)

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3315000

Table 1: Statistics on New York Times and Die Zeit corpora.

NYT	# Docs	# Timex	# Dates	# Times	# Durations	# Sets
Arts	88,822	1,258,655	947,222	134,459	159,568	18,406
Business	113,305	593,669	551,272	550	41,043	804
Sports	72,004	1,114,058	816,976	88,499	193,300	15,283
Politics	82,489	1,183,698	944,445	50,275	177,160	11,818
Die Zeit	# Docs	# Timex	# Dates	# Times	# Durations	# Sets
Arts	32,099	296,198	244,967	4,100	44,350	2,781
Business	20,572	222,577	168,215	2,023	46,420	5,919
Sports	830	6,637	4,474	204	1,420	139
Politics	31,575	281,799	216,445	4,284	56,652	4,418

For instance, a Year-granularity expression followed by one with Day-granularity form the temporal bigram [Year, Day] that is found in: "*This year*, the championships started on the *6th of May*."

Further advanced features (referred to as *Relations* in Sec. 3) are the reference direction of a temporal expression (past, present, future) with respect to the document's publication date. That is, using the above features we replaced, for instance, Day with Day-Future for the expression *tomorrow* and Year with Year-Present for *this year*. In addition, we use the realization type of the temporal expressions, i.e., assigned information if an expression occurred explicitly or relatively etc. For instance, *2001* (assuming a document publication date of 2001) would be represented as Year-Present-Explicit, while *tomorrow* as Day-Future-Relative.

Data and Preprocessing. We use the *New York Times* corpus, a collection of English daily news articles from 1987 to 2005 and the *Die Zeit* corpus which contains German articles of a weekly newspaper from 1995 to 2011. The documents from both corpora were automatically annotated with the temporal tagger HeidelTime [8].

For our experiments, we selected the newspaper categories Arts (literature and music), Business (companies), Sports (football and baseball), and Politics (elections and terrorism). The data sets were split randomly into test and training sets, with a test set size of 1,000 documents per category. All remaining documents were used for training. Table 1 provides statistics about the corpora. Note that the category information was extracted from the documents' meta data. While all categories of the English data set contain about 70,000 to 110,000 documents, the German data set has only very few sports articles, which we thus excluded in our experiments.

Classification. We use Weka [5] to train k-Nearest-Neighbours and Decision Trees with the above described features.

3 EVALUATION

A simple baseline to classify documents is to assign the same class to all documents. With fixed test sizes this results in a 25% success rate for 4 classes. As shown in Table 2(a), rather simple features like the counts of the 4 TIMEX types give far better results with up to 51.5%. A reason is probably that the distribution of expression types is quite different across categories, e.g., art documents contain many TIME-typed expressions, but only few occur in business reports.

The generalized normalized values and the temporal bigrams based on them give even better results with up to 66.3% and 64.7%, respectively. Even if the bigrams seem a little less descriptive for the classification, combining both features leads to a 66.7% success rate. The best results were achieved with the generalized normalized values extended with their temporal relation to the publication date and their realization (decision tree: 69.3%, 9-NN: 68.2%).

Table 2: Accuracy in % with 4 classes (a) and 3 classes (w/o Politics class) (b) for the New York Times corpus, and 3 classes for the Die Zeit corpus (w/o Sports class) (c).

(a)	Types	Values (V)	Bigrams	V + Bigrams	V + Relations
9-NN	51.3	64.5	63.0	66.7	68.2
Decision Tree	51.5	66.3	64.5	65.6	**69.3**
(b)	Types	Values (V)	Bigrams	V + Bigrams	V + Relations
9-NN	64.6	78.7	77.1	77.1	82.3
Decision Tree	65.7	80.8	80.1	80.4	**83.3**
(c)	Types	Values (V)	Bigrams	V + Bigrams	V + Relations
9-NN	41.4	47.3	46.0	46.0	50.4
Decision Tree	42.0	47.0	46.8	47.1	**51.4**

The decision tree classifier also outperforms the 9-NN classifier for tests with 3 classes with 83.3% for New York Times (Table 2(b)) and 51.4% for Die Zeit (Table 2(c)). However, classification of the German texts is not as good compared to their English counterparts, even though there is a similar increase for the different features. This could be explained by the fact that the New York Times training sets were three to four times larger. In addition, the weekly newspaper Die Zeit has longer and more unique-style documents compared to the daily New York Times articles, which are more standard and seem to have more prototypical temporal signatures.

4 SUMMARY AND ONGOING WORK

We presented the first experiments towards classifying documents solely based on temporal information extracted from respective texts. We have developed time-centric features and tested classifiers that achieve up to 83.3% accuracy for 3 classes and 69.3% for 4 classes. Though the results are encouraging, there is room for improvements. We currently test our approach using more classes. We plan to also switch the domain of the documents to test if similarly promising results can be achieved on privacy-sensitive data.

REFERENCES

[1] Prabal Agarwal and Jannik Strötgen. 2017. Tiwiki: Searching Wikipedia with Temporal Constraints. In *Proc. of WWW (WWW'17 Companion)*. 1595–1600. https://doi.org/10.1145/3041021.3051112

[2] Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. 2009. Clustering and Exploring Search Results Using Timeline Constructions. In *Proc. of CIKM*. 97–106. https://doi.org/10.1145/1645953.1645968

[3] Ching-man Au Yeung and Adam Jatowt. 2011. Studying How the Past is Remembered: Towards Computational History Through Large Scale Text Mining. In *Proc. of CIKM*. 1231–1240. https://doi.org/10.1145/2063576.2063755

[4] Nathanael Chambers. 2012. Labeling Documents with Timestamps: Learning from their Time Expressions. In *Proc. of ACL*. 98–106. http://dl.acm.org/citation.cfm?id=2390524.2390539

[5] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1 (Nov. 2009), 10–18. https://doi.org/10.1145/1656274.1656278

[6] Thorsten Joachims. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proc. of ECML*. 137–142. http://dl.acm.org/citation.cfm?id=645326.649721

[7] Leonardo Rocha, Fernando Mourão, Adriano Pereira, Marcos André Gonçalves, and Wagner Meira, Jr. 2008. Exploiting Temporal Contexts in Text Classification. In *Proc. of CIKM*. 243–252. https://doi.org/10.1145/1458082.1458117

[8] Jannik Strötgen and Michael Gertz. 2013. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation* 47, 2 (2013), 269–298. https://doi.org/10.1007/s10579-012-9179-y

[9] Jannik Strötgen and Michael Gertz. 2016. Domain-Sensitive Temporal Tagging. *Synthesis Lectures on Human Language Technologies* 9, 3 (2016). https://doi.org/10.2200/S00721ED1V01Y201606HLT036

[10] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Proc. of NIPS*. 649–657. http://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf