Temporal Feature Space for Text Classification

Stefano Giovanni Rizzo(✉) and Danilo Montesi

Department of Computer Science and Engineering, University of Bologna,
Mura Anteo Zamboni 7, Bologna, Italy
{stefano.rizzo8,danilo.montesi}@unibo.it
http://www.smartdata.cs.unibo.it/

Abstract. In supervised learning algorithms for text classification the text content is usually represented using the frequencies of the words it contains, ignoring their semantic and their relationships. Words within *temporal expressions* such as "*today*" or "*last February*" are particularly affected by this simplification: the same expression can have a different semantic in documents with different timestamps, while different expressions could refer to the same time. After extracting temporal expressions in documents, we model a set of temporal features derived from the time mentioned in the document, showing the relation between these features and the belonging category. We test our temporal approach on a subset of the New York Times corpus showing a significant improvement over the text-only baseline.

Keywords: Temporal features · Automatic text classification · Semantic annotation

1 Introduction

The goal of text classification tasks is to assign the category of a text document (such as an email or a web page) given the features that represent its content. Usually, a document is represented using a bag-of-word: a boolean vector with one element for each word in the document collection. In the bag-of-word representation, the feature (i.e. an element of the boolean vector) denotes the presence or the absence of each word. A text classifier, trained using these features, will estimate the category of a document given the presence or absence of the more representative words.

Intuitively, the better the features can describe documents with respect to their categories, the higher will be the accuracy of a model trained with such features. Processing the text features using NLP techniques with the goal of enhancing accuracy has a long history in Information Retrieval and Text Classification tasks. In practice, most of the attempts made using a richer, linguistic representation of text, instead of traditional word tokens, were ineffective, bringing no evidence to justify a text processing much more complex than the simple tokenization [10]. More recently however, the use of semantic annotations in addition to tokens has improved the accuracy of retrieval tasks significantly [6,9,14]. Moreover, a new

K. Ekštein and V. Matoušek (Eds.): TSD 2017, LNAI 10415, pp. 362–370, 2017.
DOI: 10.1007/978-3-319-64206-2_41

trend is emerging which gives particular attention on the temporal dimension of text [4], and for which temporal semantic annotation and extraction is a crucial step [1,3].

The particular interest towards content-level temporal information resides in its great variance in expressing the same object, because synonymy and polysemy relations in temporal expressions (timexes) are more complex than in other named entities:

- Synonyms of absolute timexes: any absolute mention of a time interval such as "*4 October 2016*", can have a number of variations for each mentionable time interval (e.g. "*4/10/2016*", "*4 October 2016*", "*the fourth of October 2016*" etc.).
- Synonyms of relative timexes: the same moment in time can be mentioned using a relative time expression such as "*tomorrow*" or "*one year ago*", depending on the time of writing. What was "*today*" for a philosopher of the ancient Greece becomes "*two thousand years ago*" in the present time.
- Hypernyms of relative timexes: the same time expression can refer to any interval, depending on the time of writing: "*next year*" could refer to 1950 if written in 1949, or could refer to 2017 if written in 2016.

Improving text categorization by means of semantic annotations of named entities has been considered in the past [2], however to the best of our knowledge no work has been done to exploit content-level time for text categorization.

In this paper we propose a temporal features space, in addition to the traditional text features space, to improve text classification tasks such as topic categorization or new event detection. The set of novel temporal features for documents, derived from time mentioned in text, captures the temporal peculiarities of the documents related to their category in a low-dimensional representation. These features take in consideration how much time is mentioned in a document, the central time of the narrated events, and the span of the intervals cited.

After formally defining how these features are built, we show the results of ANOVA (analysis of variance) to assure correlation between temporal features and categories on the New York Times dataset, a well-known corpus of manually categorized documents. Finally, we evaluate how much accuracy improvement is obtained using both temporal and text features sets.

2 Temporal Feature Space

Each document, in its textual content, cites a number of absolute and relative dates (*content-level time*). For instance, a certain document can contain a temporal expression such as "On *2016 Christmas eve*" referring to the absolute date 2016-12-24. The same document could also contain a relative temporal expression such as "the match we watched *yesterday*", referring to a relative date, which depends on the creation time of the document (a timestamp known as document creation time or DCT). All the temporal expressions, both absolute

and relative, can be annotated and normalized into exact timestamp intervals using a temporal annotator, such as Heideltime [13]. We define this set of all the mentioned intervals as the *temporal scope* of the document.

The temporal features we define are all derived from the temporal scope of documents. From this set of intervals, we extract different time features that are able to finely describe the document characteristic in the temporal space, without using a plain *bag-of-chronons* that would require a very high-dimensionality representation:

1. **Temporality feature:** some texts are more time-related than others (e.g. news article vs philosophy argument). In the same way, documents belonging to different categories can have a significantly different *temporality*. Temporality is an indicator of how much time there is in a document.
2. **Focus and mean time features:** these two features denote the central scope of the intervals mentioned in the text, with two different notions on what is the central time window of the narrated events.
3. **Interval size feature:** depending on the topic, the mentioned intervals can be short, such as one day, or longer such as years and millenniums. The interval size considers the span's length of the mentioned intervals.

2.1 Temporality

We define the temporality of a document as the number of timexes in a document. Despite its simplicity, this feature captures a property that can strongly discern some topics and categories. This is due to the fact that the subjects of some categories rely on many time mentions, while others hardly make use of time in their narrative.

Definition 1 (Temporality). *Given the temporal scope T_D of a document as the set of all the mentioned intervals in its content, the* ***temporality*** *is the cardinality of T_D.*

$$time_{temporality}(T_D) = |T_D| \tag{1}$$

2.2 Mean Time and Focus Time

The set of time expressions in a document often revolves around a central time window, such as the time of the main event described. We provide two different central time definitions: mean time and focus time.

Definition 2 (Mean time window). *Given the temporal scope of a document as the set of all the mentioned intervals in its content, the* ***mean time window*** *is an interval $[t_s, t_e]$ where t_s is the mean of all the start times in the temporal scope and t_e is the mean of all the end times in the temporal scope.*

$$time_{window}(T_D) = [\frac{1}{|T_D|} \sum_{t \in T_D} t_s, \frac{1}{|T_D|} \sum_{t \in T_D} t_e] \tag{2}$$

The mean time window, aggregating all the mentioned intervals, gives a rich information on the time extent of the document. However, averaging the intervals can lose a very crucial information, which is what the "focus" of the document is.

Different works in literature have different conceptions on what the focus time of a document is. For Strotgen et al. [12] the focus time is the most frequent time in a document, while in more complex approaches [8] the focus time is the one with which the document's terms are mostly associated in the corpus. Following the former notion of focus time [12], we define our focus time as the mode of the frequency distribution, that is, the interval which is most frequently mentioned in the document.

Definition 3 (Focus time). *Given the temporal scope of a document as the set of all the mentioned intervals in its content, the* ***focus time*** *is an interval* $[m_s, m_e]$ *where* m_s *is the mode of all the start times in the temporal scope and* m_e *is the mode of all the end times in the temporal scope.*

$$time_{focus}(T_D) = [mode(t_s), mode(t_e)] \tag{3}$$

In order to illustrate how well the focus time can approximate the time of a document, and the difference between the focus time and the mean time window, we picked two very different documents.

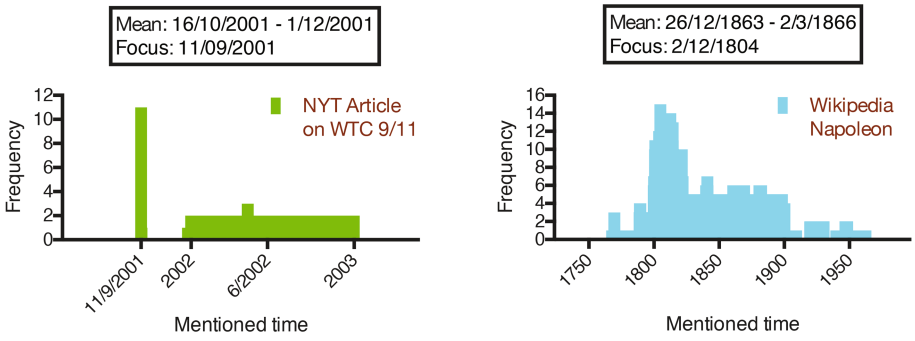

Fig. 1. Temporal scope for two documents: a New York Times article on the WTC terrorist attack and the Wikipedia main article on Napoleon. Top frame shows Mean time and Focus time.

In Fig. 1 we represent the content time of the two different documents as a frequency distribution of each interval. On the left, a New York Times article on the WTC 2001 attack, written in 2002, shows a peak on the day of the attack, but it also shows other mentioned intervals about events happened after the attack, along the year 2002. For this reason, the mean time window for this article is the period from 16/10/2001 to 1/12/2001, while the focus time identifies only the main event time as the 11th of September 2001. The same results are obtained on a totally different kind of text document: the Wikipedia article on Napoleon

Bonaparte. The focus time is 2 December 1804, the date of his incoronation, while the mean takes into account all the related events.

2.3 Interval Size

The temporal expressions found in a document, once normalized to time intervals, can have a different size depending on the span of the cited time. For instance, in the gregorian calendar, this time span can be of 7 days if a week is mentioned, 28 to 31 days if a month is mentioned, 365 days for the year and so on. Moreover, there can be found a smaller percentage of irregular intervals, in temporal expressions such as "*I will train for 10 days*" or "*The Great War lasted from 28 July 1914 to 11 November 1918*". Put together, all these intervals compose a set that can be very diversified, but can also follow some patterns depending on the topic of a document and, therefore, on its category.

Definition 4 (Intervals size). *The intervals size of a document is the mean size of all the intervals of its temporal scope.*

$$time_{size} = \frac{\sum_{x \in T_D} x_e - x_s}{|T_D|} \tag{4}$$

This feature, although simple in its definition, is valuable in discriminating documents from different categories, as we show in the next section.

3 Experimental Results

We evaluate the proposed features on a random subset[1] of 75 thousand documents from the The New York Times Annotated Corpus[2], which is the most used corpus for temporal related tasks [1,11] because of its temporal richness both in content (we extracted 15 million temporal expressions over 1.8 million documents) and in production time variance (it spans over 20 years of articles). Timexes have been identified and normalized using Heideltime [13], considered the state of the art tagger able to recognize even BC period dates. The category annotation of New York Times articles is provided by the New York Times Newsroom, the New York Times Indexing Service and the online production staff at nytimes.com.

After annotating all the temporal expressions in the corpus and extracting the normalized time intervals, we randomly sampled 5,000 articles for each one of the 15 most occurrent online section categories: *Arts*, *Business*, *Magazine*, *New York and Region*, *Obituaries*, *Opinion*, *Paid Death Notices*, *Real Estate*, *Sports*, *Style*, *Technology*, *Travel*, *U.S.*, *Week in Review*, *World*.

[1] Dataset with precomputed features available at https://smartdata.cs.unibo.it/data/TFTC/.

[2] NYT Corpus available at https://catalog.ldc.upenn.edu/LDC2008T19.

3.1 Significance Study

The ANOVA test (analysis of variance) is a statistical model to analyze the difference for a specific variable (feature) over a set of groups (categories). The ANOVA test has been widely used for feature selection because it gives a measure of the *reliability* of a feature [7]. All the proposed features have significantly different averages across different categories ($\texttt{p-value} \ll 0.01$), meaning that the difference between the mean values of the categories, for each defined feature, is not due to chance.

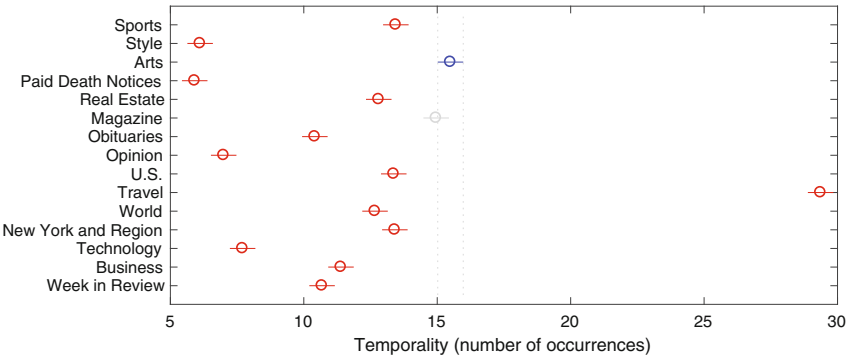

Fig. 2. Mean (circle) and variance (line) of **temporality** for each category in NYT Section dataset. (Color figure online)

In Fig. 2 we show the mean temporality for each category and the variance within the same category. Circle points denote the mean temporality for the category, while horizontal lines over each point denote the extension of the variance.

By looking at the variance lines overlaps, it is possible to visually identify categories with a significantly different temporality: if the variance line for two or more categories overlaps, these categories do not significantly differ, thus a classifier trained only on this feature will not be able to discriminate between them. As an example, in Fig. 2 we highlight in blue the *Arts* category, while in red are shown 13 categories that significantly differs from *Arts*. In gray, the *Magazine* category is the only category which cannot be distinguished from *Arts* with significant confidence.

Among the fifteen categories, the mean time is less scattered than the other features, as shown in Fig. 3. Despite this most categories are significantly different from each other. The highlighted examples in blue is the category *Real Estate*: the average of its mean time is totally indistinguishable from the category *Travel*, however it is well distinguishable from the other 13 categories.

The means of interval sizes in Fig. 4 are also quite diverse among categories. The category *U.S.*, highlighted in blue, is an unlucky example: the sizes of its mentioned intervals are similar to *World*, *New York and Region* and *Opinion*, while significantly different from the other 11 categories.

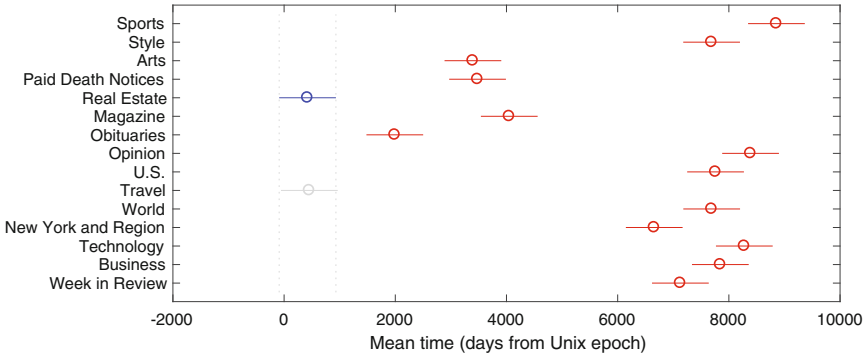

Fig. 3. Mean (circle) and variance (line) of **mean time** for each category in NYT Section dataset. (Color figure online)

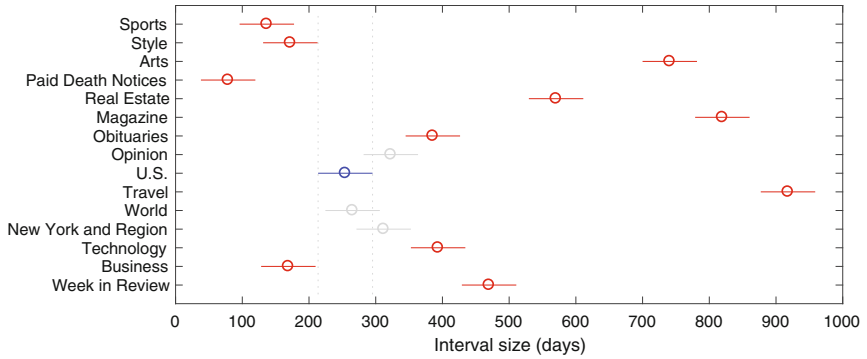

Fig. 4. Mean (circle) and variance (line) of **interval size** for each category in NYT Section dataset. (Color figure online)

It is noteworthy that, for each pair of categories in this corpus there exists at least one of the defined features for which they significantly differ.

3.2 Classification Accuracy

To test the improvement in accuracy we combine two simple kNN (k Nearest Neighbor) classifiers: one trained on the text features (*tf.idf* vector), the other trained using the novel temporal features. We set k as 35 for both classifiers since this is a known good choice to yield stable effectiveness [16]. Once we evaluate the accuracy for the single classifier, we use this accuracy as the weight in the linear combination of the two classifier [15]. We obtain an F1 macro-averaged accuracy of 0.247 using the temporal features only and 0.681 using the traditional *tf.idf* features, while combining both temporal and text features we obtain 0.694.

We further test the improvement using a state-of-the-art classifier, the tree boosting model XGBoost [5], on the union set of both feature spaces, this time

using a single classifier. In this setting we obtain 0.417 using only time features, 0.852 using text features and 0.864 using both.

Applying the t-test on the predicted categories of all tests it results that the improvement obtained adding the temporal features is not due to chance ($\texttt{p-value} \ll 0.01$).

4 Conclusion

In this paper we have shown how the time mentioned in documents is category-dependent, thus can be exploited to recognize the category of documents with higher accuracy. We defined a set of time features to synthesize the temporal scopes of documents by their centrality, extent and size. Analyzing the variance of these features among different categories, it results that most categories are far apart from a temporal point of view while smaller groups of categories share similar temporal aspects. The experimental evaluation confirms that while using solely the proposed temporal features does not lead to adequate accuracy, their combination with traditional terms-based features yields a significant improvement over the text-only baseline.

References

1. Berberich, K., Bedathur, S., Alonso, O., Weikum, G.: A language modeling approach for temporal information needs. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 13–25. Springer, Heidelberg (2010). doi:10.1007/978-3-642-12275-0_5
2. Bloehdorn, S., Hotho, A.: Boosting for text classification with semantic features. In: Mobasher, B., Nasraoui, O., Liu, B., Masand, B. (eds.) WebKDD 2004. LNCS, vol. 3932, pp. 149–166. Springer, Heidelberg (2006). doi:10.1007/11899402_10
3. Brucato, M., Montesi, D.: Metric spaces for temporal information retrieval. In: Rijke, M., Kenter, T., Vries, A.P., Zhai, C.X., Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 385–397. Springer, Cham (2014). doi:10.1007/978-3-319-06028-6_32
4. Campos, R., Dias, G., Jorge, A.M., Jatowt, A.: Survey of temporal information retrieval and related applications. ACM Comput. Surv. (CSUR) **47**(2), 15 (2015)
5. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM (2016)
6. Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., Motta, E.: Semantically enhanced information retrieval: an ontology-based approach. Web Semant. Sci. Serv. Agents World Wide Web **9**(4), 434–452 (2011). JWS special issue on Semantic Search
7. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
8. Jatowt, A., Au Yeung, C.M., Tanaka, K.: Estimating document focus time. In: Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management, pp. 2273–2278. ACM (2013)

9. Kara, S., Alan, Ö., Sabuncu, O., Akpinar, S., Cicekli, N.K., Alpaslan, F.N.: An ontology-based retrieval system using semantic indexing. Inf. Syst. **37**(4), 294–305 (2012). Semantic Web Data Management
10. Moschitti, A., Basili, R.: Complex linguistic features for text classification: a comprehensive study. In: McDonald, S., Tait, J. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 181–196. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24752-4_14
11. Radinsky, K., Agichtein, E., Gabrilovich, E., Markovitch, S.: A word at a time: computing word relatedness using temporal semantic analysis. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, pp. 337–346. ACM, New York (2011)
12. Strötgen, J., Alonso, O., Gertz, M.: Identification of top relevant temporal expressions in documents. In: Proceedings of the 2nd Temporal Web Analytics Workshop, pp. 33–40. ACM (2012)
13. Strötgen, J., Gertz, M.: Heideltime: High quality rule-based extraction and normalization of temporal expressions. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 321–324. Association for Computational Linguistics (2010)
14. Vallet, D., Fernández, M., Castells, P.: An ontology-based information retrieval model. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 455–470. Springer, Heidelberg (2005). doi:10.1007/11431053_31
15. Wu, S., Crestani, F.: Data fusion with estimated weights. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, pp. 648–651. ACM (2002)
16. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–49. ACM (1999)