Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 1070 – 1085
© TÜBİTAK
doi:10.3906/elk-1907-10

Estimating spatiotemporal focus of documents using entropy with PMI

Damla YAŞAR, Selma TEKİR*
Department of Computer Engineering, Faculty of Engineering, İzmir Institute of Technology, İzmir, Turkey

Received: 02.07.2019 • **Accepted/Published Online:** 12.12.2019 • **Final Version:** 28.03.2020

Abstract: Many text documents are spatiotemporal in nature, i.e. contents of a document can be mapped to a specific time period or location. For example, a news article about the French Revolution can be mapped to year 1789 as time and France as place. Identifying this time period and location associated with the document can be useful for various downstream applications such as document reasoning or spatiotemporal information retrieval. In this paper, temporal entropy with pointwise mutual information (PMI) is proposed to estimate the temporal focus of a document. PMI is used to measure the association of words with time expressions. Moreover, a word's temporal entropy is considered as a weight to its association with a time point and a single time point with the highest overall score is chosen as the focus time of a document. The proposed method is generic in the sense that it can also be applied for spatial focus estimation of documents. In the case of spatial entropy with PMI, PMI is used to calculate the association between words and place entities. The effectiveness of our proposed methods for spatiotemporal focus estimation is evaluated on diverse datasets of text documents. The experimental evaluation confirms the superiority of our proposed temporal and spatial focus estimation methods.

Key words: Document analysis, spatiotemporal focus estimation, temporal entropy, spatial entropy, pointwise mutual information

1. Introduction

Mapping contents of a document to a specific time period or location is important for document understanding. Location and time expressions can be used as clues to predict the spatiotemporal focus of unstructured text. A key approach in document focus time estimation [1] relies on the association of words with time expressions in documents and the time expression that has a distinctive association with terms in documents is selected as the target. Words' contributions to the overall document-time association score are weighted with respect to how strongly associated they are with time points. The scheme depends on the extraction of time expressions. As place entities can be extracted in a similar way, the proposed approach is also applicable to focus place estimation of documents.

Word association calculations in the existing literature for document focus time estimation are performed using the Jaccard coefficient, which measures the difference of words' cooccurrences from their union event of occurrences. Pointwise mutual information (PMI) is a good measure of association between words in that it is based on probabilities rather than raw frequencies. PMI measures the distinction of joint probability from that of individual probabilities, assuming independence. Thus, we propose to use PMI rather than the Jaccard coefficient in those association calculations. Our results confirm the utility of using PMI as a term association measure in document spatiotemporal estimation.

*Correspondence: selmatekir@iyte.edu.tr

In order to give the intuitive idea behind PMI usage over the Jaccard coefficient as the word association measure, an example paragraph is provided from the ancient Bar Kokhba Revolt document in the Ancient Times corpus [2]. The paragraph contains two time expressions, 132–136 and 115–117, respectively. For the given paragraph, while the Jaccard coefficient points to 117 as the temporal estimation result, PMI estimates it as 136, which is the ground truth as the paragraph narrates the Bar Kokhba Revolt that took place between 132 and 136. PMI's focus on the first time expression rather than the second can be explained by the fact that document focus time estimation relies on how strongly associated a time point is with the words in the document. For the year 136, PMI returns substantially higher association scores with the words revolt, rebellion, against, and climax when compared to those of the year 117. However, the Jaccard coefficient cannot make a distinction between 136 and 117, and it fails to give the correct focus time estimation as 136.

Our contributions can be outlined as follows:

- We extend the existing work on temporal entropy with the Jaccard coefficient [1] by the use of PMI rather than the Jaccard coefficient as the word association measure. We call our approach temporal entropy with PMI.
- Time and place are intrinsically different from each other. We can represent time points on a line, whereas a geodesic grid is used to display place points. Due to these distinct representations, we use different metrics to calculate distance among time and place points, respectively. Also, time and place estimation approaches differ in that in time estimation, language models are learned for certain time intervals, while in the case of place estimation, learning is performed on a grid cell basis. We propose spatial entropy with PMI in order to estimate the focus place of documents. Despite the inherent differences between time and location estimation, our approach with PMI works in both domains.
- We expanded the experimental basis for testing document time estimation models by adding two new datasets that are developed from the History World and Britannica Biographies datasets, respectively.

In the remaining part of the paper, first in Section 2, literature on temporal and spatial estimation of documents is introduced. After that, in Section 3, methods and materials are presented. In Section 3, initially preprocessing of text documents and the knowledge base for association calculation are explained. Then, first for the temporal estimation and then for the spatial one, the proposed approaches are described. At the end of the section, datasets and results of the experimental evaluation are given. Finally, in Section 4, a discussion about the results is given, and in Section 5, the paper is concluded with some remarks and possible future directions.

2. Related work

Most of the work on document dating has been carried out by information extraction and information management communities interested in dating documents with unknown origins. Typical approaches to dating documents are based on the change of language over time and use temporal language models. De Jong et al.'s work [3] is among the first studies to automatically timestamp documents. Unigram language models are learned for certain time intervals and articles are scored according to log-likelihood ratio scores. Kanhabua and Norvag [4, 5] extended the same approach with POS tags, collocations, and tf-idf features.

Kumar et al. [6] focused on dating Gutenberg short stories. Unigram language models are learned as in previous studies, but unlike previous ones, the KL-divergence value between the document and the language model of a time interval is measured. Kumar et al. [7] developed language models of years based on Wikipedia biography pages in order to estimate the focus time of documents.

Dalli [8] calculated the probability distributions over different time intervals (months and years) for each observed period. The focus is on finding words whose time distribution has high standard deviation weighted by frequency scores. In [9], Chambers extended previous approaches by focusing on language structure and absolute time expressions. Niculae et al. [10] provided a new approach to the task of classifying temporal texts by combining text order and probability in automatic dating of historical texts. In addition, a series of research articles have been published based on a heuristic method to automatically create or verify the temporal metadata of historical texts [11, 12].

In [1], Jatowt et al. used collocations in order to determine the strongly related words with time expressions and estimate the time period of the document relying on them.

The closest work to ours is the work of [1]. Distinct from that work, we use PMI as the association score to measure how strongly a word and a time expression are related. We use Wikipedia articles as the underlying knowledge base to compute our association scores.

Many text documents from all kinds of different domains are said to be related to some geographic content. Therefore, there are a great deal of works that have been published on text-based geotagging and geocoding. Geotagging or geocoding is a procedure to add geographical metadata that map onto a whole document according to its textual contents. A geotag can define the longitude and latitude of a tagged text document or can define the location place name or regional identifier. Melo and Martins [13] classified different geotagging approaches and highlighted the rise of discriminative classification models that were proposed recently.

The first work that proposed a geotagging mechanism for text documents was that by Woodruff and Flaunt [14]. Their proposed system, which is named GIPSY, geotags textual documents in order to help indexing and retrieval in document search.

Martins et al. [15] developed methods in order to extract semantic spatiotemporal information from text using simple text mining methods on a gazetteer. The proposed system is capable of displaying information over maps and timelines. Han et al. [16] concentrated on retrieving location indicative words by feature selection and observed that the reduced feature set increases geolocation accuracy.

Zhang and Gelernter et al. [17] proposed a supervised machine learning model to evaluate the features of a gazetteer and tags of tweets to geotag a location.

Laere et al. [18] developed probabilistic language models trained on Flickr and Twitter to predict the coordinates of Wikipedia articles. Wing and Baldridge [19] demonstrated the effectiveness of using logistic regression models on a hierarchy of nodes in a geodesic grid, unlike previous works that ignored the natural hierarchy of cells in such grids. Priedhorsky et al. [20] proposed a content-based approach to estimate the location of tweets using a variant of Gaussian mixture models.

Li et al. [21] proposed a three-step method for the purpose of identifying top-k locations of a microblog and top-k locations of a user. They proposed a global location identification method, named GLITTER, which puts microblogs of a user into a tree. GLITTER extracts candidate locations from tree nodes, aggregates these candidate locations, and identifies top-k locations of the user. As the final step, the system refines the candidate locations using these top-k user locations and computes the top-k locations of each microblog.

Laere et al. [22] proposed two classes of term selection techniques based on statistical methods of spatial terms. First, to implement the idea of spatial smoothing of term occurrences, they investigated the use of kernel density estimation (KDE) to model each term as a two-dimensional probability distribution over the surface of the Earth. The second class of term selection methods they considered is based on Ripley's K statistic, which measures the deviation of a point set from spatial homogeneity.

Research on automatically geolocating social media users has conventionally been based on the text content of posts from a given user or the social network of the user, with very little crossover between the two. Rahimi et al. [23] brought the two threads of research together in first proposing a text-based method based on adaptive grids, followed by a hybrid, network, and text-based method.

Grid-based approaches suffer from data sparsity if one wants to improve classification accuracy by moving to smaller cell sizes. Hulden et al. [24] investigated an enhancement of common methods for determining the geographic point of origin of a text document by kernel density estimation.

Brunsting et al. [25] presented an algorithm for location tagging of textual documents. They used a state-of-the-art part-of-speech tagger and named entity recognizer to find blocks of text referring to locations. Afterwards, a knowledge base named OpenStreetMap[1] was put to use in order to find a list of possible locations. Finally, one location is chosen by assigning distance-based scores to each location and choosing the highest score.

Rodrigues et al. [26] proposed a probabilistic approach that jointly models geographical labels and Twitter texts of users organized in the form of a graph representing the friendship network. They used the Markov random field probability model to represent the network and learning is carried out through a Markov chain Monte Carlo simulation technique to approximate the posterior probability distribution of the missing geographical labels.

Kordopatis-Zilos et al. [27] presented an approach for estimating locations using text annotations based on refined language models that are learned from massive corpora of social media annotations. They also explored the impact of different feature selection and weighting techniques on the performance of the approach.

Research on geotagging and georeferencing focuses on two main methods: gazetteer-based methods and language modeling-based methods. Our work is different from these works in that we use local context-based association (PMI) of words with location names to estimate the location of a document. As a result of our literature research, a study using the aforementioned method could not be found. Another difference of our work is that our proposed method is shown to perform well on both social media and Wikipedia data.

3. Methods and materials

A general document can contain several time and place expressions. Some of them may be unrelated to the document content. The objective of the present research is to compute the focus time and focus place of a document with respect to its content.

We propose temporal entropy with PMI that automatically categorizes documents by their temporal focus. The proposed method utilizes statistical knowledge from external resources, Wikipedia in our case. The statistical knowledge from the Wikipedia knowledge base is represented through the use of word association measures.

3.1. Information extraction

In order to estimate the spatiotemporal focus of a text document, we need to extract time expressions and place entities first. Then we calculate word-time and word-place association scores to use in our estimation model.

[1] OpenStreetMap (2013). OpenStreetMap [online]. Website https://www.openstreetmap.org [accessed 10 January 2019].

In order to extract time expressions, we use the temporal tagger HeidelTime [28], the first multilingual, cross-domain temporal tagger for the full task of temporal tagging, i.e. performing the extraction and the normalization of temporal expressions. HeidelTime uses the TimeML annotation standard with TIMEX3 tags for temporal expressions since it is the most recent standard. The TIMEX3 tag is primarily used to mark up explicit temporal expressions, such as times, dates, durations, etc. It is modeled on Setzer's TIMEX tag [29], as well as the TIDES TIMEX2 tag [30]. HeidelTime[2] is publicly available and it has already been used in several studies by other research groups.

Named entity recognition (NER) is one of the information extraction tasks that aims to locate and classify named entities in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. The term "Named Entity" was coined at the 6th Message Understanding Conference (MUC) [31]. We use the Stanford NER (Named Entity Recognizer) to extract place entities from a given text. NLTK provided an interface for Stanford NER[3] in Python.

3.2. Preparing the knowledge base

In order to calculate word-time point association scores, an adequate amount of temporal references needs to be collected. This will serve as a knowledge base while computing word-time point association scores in estimating document focus time. A 2016 dump of the English Wikipedia corpus is used to create a knowledge base to ground our word-time association scores. The corpus has 112,409,626 tokens before stop words are removed. After removing the stop words, the corpus is left with 70,283,848 tokens.

3.3. Temporal focus estimation

3.3.1. Temporal entropy with PMI

In estimating the temporal focus of documents, we use the association of words with time points. Thus, as a preprocessing step, we extract the time expressions contained in the document by using the HeidelTime temporal tagger [28]. If the document does not contain any time expressions, we estimate the temporal focus by calculating its association with all time points in year granularity between B.C. 2000 and A.C. 2000.

When it comes to calculation of word-time point associations, in the literature, the Jaccard coefficient [32] is used for this purpose [1]. We propose to use PMI [33] as a word association measure in calculating a document's temporal focus. In PMI (Equation 1), the joint probability of two words are normalized by the product of their marginals. Normally PMI can take negative values. However, in our experiments, since the stop words in the documents are removed in the preprocessing phase, we do not get any negative PMI values. The Jaccard coefficient (Equation 2), on the other hand, is the ratio of two words' cooccurrences to the difference of the sum of their marginal occurrences and cooccurrences.

$$A_{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i) \times p(w_j)} \tag{1}$$

$$A_{jac}(w_i, w_j) = \frac{c(w_i, w_j)}{c(w_i) + c(w_j) - c(w_i, w_j)} \tag{2}$$

[2] HeidelTime (2015). HeidelTime [online]. Website https://github.com/HeidelTime/heideltime [accessed 15 April 2018].

[3] NLTK (2019). NLTK 3.4.5 documentation nltk.tag package [online]. Website http://www.nltk.org/api/nltk.tag.html [accessed 20 January 2019].

In order to estimate the focus time of documents, we use the same hypothesis, "word w has a high association score with time point t if many words that cooccur with w are also strongly associated with t" in [1].

A word w_i's association with time point t is calculated as a weighted sum of all the terms' association with t, where the square of w_i's association with each vocabulary term is used as the weight (Eq. (3)).

$$A_{time}(w_i, t) = \frac{1}{|v|} \sum_{j=1}^{|v|} A(w_i, w_j)^2 \times A(w_j, t) \tag{3}$$

Furthermore, the association score of a word w_i with time point t is normalized after dividing this value by the geometric mean of the association scores of all other words with the time point t.

Not all words have equal discriminative capabilities. A word has high discriminative capability for determining document focus time if it has strong association with only a few time points while having weak association with the rest of them.

Temporal entropy indicates a word's discriminative capability by scoring high when a word's association with time shows a nonuniform probability distribution. Therefore, in order to score words regarding their discriminative capabilities, temporal entropies of words are computed.

Assuming that $P_w(t_i)$ represents the probability of a word w to be associated with time point t_i, we define the temporal entropy ω_w^{temE} of word w as follows (Eq. (4) and (5)):

$$temE_w = -\sum_i P_w(t_i) \times \ln P_w(t_i) \tag{4}$$

$$\omega_w^{temE} = max_j(temE_j) - temE_w \tag{5}$$

As given in Eq. (5), a target word's temporal entropy is calculated by subtracting its temporal entropy from the maximum temporal entropy of a word in the vocabulary.

As the final step, we calculate document-time association scores using the intuition given in [1]: "The more words strongly associated with time point t contained in a document d, the more it is likely that t belongs to the focus time of d." The formulation is given in Eq. (6).

$$S_U(d, t) = \frac{1}{|d|} \sum_{w \epsilon d} \omega_w^{temE} \times A_{time}(w, t) \tag{6}$$

In Eq. (6), every word's association score with a time point t is multiplied by that word's temporal entropy and summed over all the vocabulary terms, and then a document length-based normalization is applied.

In order to determine document focus time, a single time with the highest association score is chosen. We denote this as $t_{foc}^{ins}(d)$ in Eq. (7):

$$t_{foc}^{ins}(d) = argmax_t S_U(d, t) \tag{7}$$

3.3.2. Datasets

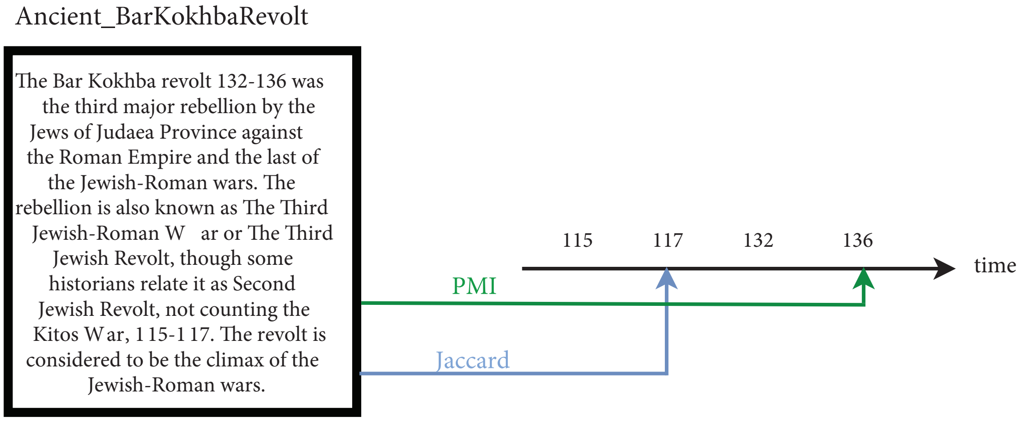
In order to test our proposed document time estimation models, we use four different datasets, namely Ancient-Times, WikiWars, HistoryWorld, and Britannica Biographies.

AncientTimes was developed by Strotgen and Gertz [2] in the context of a study on temporal tagging of texts about history. The multilingual AncientTimes corpus contains Wikipedia articles covering different time periods, and all occurring temporal expressions are manually annotated.

WikiWars presented by Mazur and Dale [34] is a corpus of 22 documents sourced from English Wikipedia. Most of the documents are war descriptions.

In order to test our focus time estimations, we developed the HistoryWorld corpus out of the History World dataset.[4] The corpus is composed of 68 documents in English. We create a text document for each entry in the History World dataset. All of the documents are about important historical events and people whose lifespan is known. Therefore, it is easy to map each document to a time point according to its content. The corpus that we built is bigger in size than WikiWars and AncientTimes. Its subject diversity is also higher when compared to those two corpora.

Although the datasets that were demonstrated previously are sufficient in having enough time and place expressions, the number of documents that they contain is small. In the light of this requirement, we choose Britannica Biographies [5] to create a dataset of 400 documents about the lifetime of famous historical figures. There are eight categories that include persons known for their achievements in fields of arts and visual design, education, entertainment, history and society, literature, philosophy and religion, sciences, and sports. The documents are divided into six time intervals for each category: 500–1 BC, 0–499, 500–999, 1000–1499, 1500–1899, and 1900–present.



The Bar Kokhba revolt 132-136 was the third major rebellion by the Jews of Judaea Province against the Roman Empire and the last of the Jewish-Roman wars. The rebellion is also known as The Third Jewish-Roman W ar or The Third Jewish Revolt, though some historians relate it as Second Jewish Revolt, not counting the Kitos W ar, 115-117. The revolt is considered to be the climax of the Jewish-Roman wars.

Figure 1. Temporal estimation results for Jaccard coefficient and PMI for the first paragraph of the ancient Bar Kokhba Revolt document.

3.3.3. Results

To evaluate our results, we calculate the average error of the focus time estimations that fall outside the time interval of the document. For example, in Britannica Biographies, we look at whether our calculated focus time is within that person's life span (birth–death). If it is, then the error is zero. If it falls outside the interval, then we measure the distance between our estimation and the closest boundary point of the true time interval. That distance is our error. Finally, we take the average of all errors. The average focus time estimation error results of the models on test datasets are given in Table 1.

[4] Gascoigne B. (2001). HistoryWorld [online]. Website http://www.historyworld.net [accessed 15 May 2018].

[5] Britannica Biographies [online]. Website https://www.britannica.com/biographies [accessed: 18 February 2018].

Table 1. Average error of the focus time estimations of temporal entropy with Jaccard coefficient and temporal entropy with PMI after taking geometric mean of the association scores.

Dataset group	Temporal entropy with Jaccard coefficient	Temporal entropy with PMI
Ancient Times	0.51 years	0.17 years
WikiWars	1.87 years	4.45 years
History World	11.37 years	10.44 years
Britannica Biographies	14.34 years	12.05 years

As seen from the average error results, Temporal entropy with PMI has superior results with the exception of WikiWars. When we analyze the individual document scores in WikiWars, there are a few outlier scores in the case of PMI due to sentences referring to a cause in the past or including a recent mention of that war. For instance, Soviet deployment in Afghanistan took place in 1979. However, the WikiWar document describing this event includes a 1998 interview reference in which Brzezinski comments on this war event.

Besides average error, we measure the accuracy of our proposed method using precision, recall, and f-measure. It can be observed in Table 2 that temporal entropy with PMI achieves the highest accuracy.

The comparison of the proposed method with the state-of-the-art can be found in Table 3. Unfortunately, original state-of-the-art datasets [1] are not available. Therefore, test datasets are created by accessing the same web sources and following the steps described by Morbidoni and Cucchiarelli [35]. We collect paragraphs from the following web sites reporting main events related to five countries. The web sites we consider are the same as in [1]: History Orb,[6] History World,[7] BBC Timelines,[8] and Infoplease.[9] In order to reproduce a dataset as similar as possible to the dataset described in [1], we use the Wayback Machine[10] to access the snapshot of the websites recorded in January 2015, which is the date that was reported to create the web dataset. The Web Dataset contains 1007 documents referring to events in the time span of 1900–2015. As shown, our proposed method outperforms state-of-the-art methods with respect to average error.

Table 2. Temporal estimation accuracy results on four different datasets after taking geometric mean of the association scores.

Dataset group	Temporal entropy with Jaccard coefficient			Temporal entropy with PMI		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Ancient Times	0.72	0.72	0.72	0.98	0.98	0.98
WikiWars	0.77	0.78	0.78	0.85	0.86	0.86
History World	0.84	0.84	0.84	0.86	0.87	0.87
Britannica Biographies	0.85	0.84	0.85	0.91	0.91	0.91

3.4. Spatial focus estimation

3.4.1. Spatial entropy with PMI

Before starting to estimate the focus place of a document, first we extract the place entities contained in the document by using the Stanford NER Tagger. Then word-place entity associations are calculated. As for

[6] HistoryOrb (2000). HistoryOrb [online] Website http://www.historyorb.com/ [accessed 15 May 2018].

[7] Gascoigne B. (2001). HistoryWorld [online]. Website http://www.historyworld.net [accessed 15 May 2018].

[8] BBC (2018). BBC Programmes History [online] Website http://www.bbc.co.uk/history [accessed 20 May 2018].

[9] Infoplease (2017). Infoplease [online] Website http://www.infoplease.com/ [accessed 20 April 2018].

[10] Internet Archive (1996). Internet Archive [online] Website http://archive.org/web/ [accessed 10 May 2018].

Table 3. Average error of proposed methods vs. state-of-the-art approaches.

Method	Avg. error (years)
Jatowt et al. [1]	20.2
Morbidoni and Cucchiarelli [35]	15.7
Proposed method (temporal entropy with PMI)	12.73

documents without any explicit place mentions, we use GeoNames,[11] which is a gazetteer containing eleven million place names that are available for download free of charge. We just calculate spatial focus estimation values for each element in the GeoNames gazetteer and take the element with the maximum score as the spatial focus of the document.

In order to estimate the focus place of documents, we use the hypothesis that "word w has a high association score with place point p if many words that cooccur with w are also strongly associated with p."

A word w_i's association with place point p is computed as given in Eq. (8). PMI is used as the word association measure. The equation is the same as Eq. (3), except that here associations are with respect to place points instead of time points.

$$A_{place}(w_i, p) = \frac{1}{|v|} \sum_{j=1}^{|v|} A(w_i, w_j)^2 \times A(w_j, p) \tag{8}$$

Furthermore, the association score of a word w_i with place point p is normalized after dividing this value by the geometric mean of the association scores of all other words with the place point p.

Not all words have equal discriminative capabilities. A word has high discriminative capability for determining document focus place if it has strong association with only a few place points while having weak association with the rest of the places.

To score words regarding their discriminative capabilities, we compute the spatial entropy of words.

Spatial entropy calculation is performed by replacing temporal variables by their spatial counterparts in Eq. (4) and (5).

As the final step, we calculate document-place association scores using the intuition of "the more words strongly associated with a place point p contained in a document d, the more it is likely that p belongs to the focus place of d." The formulation is given in Eq. (9).

$$S_U(d, p) = \frac{1}{|d|} \sum_{w \epsilon d} \omega_w^{spatialE} \times A_{place}(w, p) \tag{9}$$

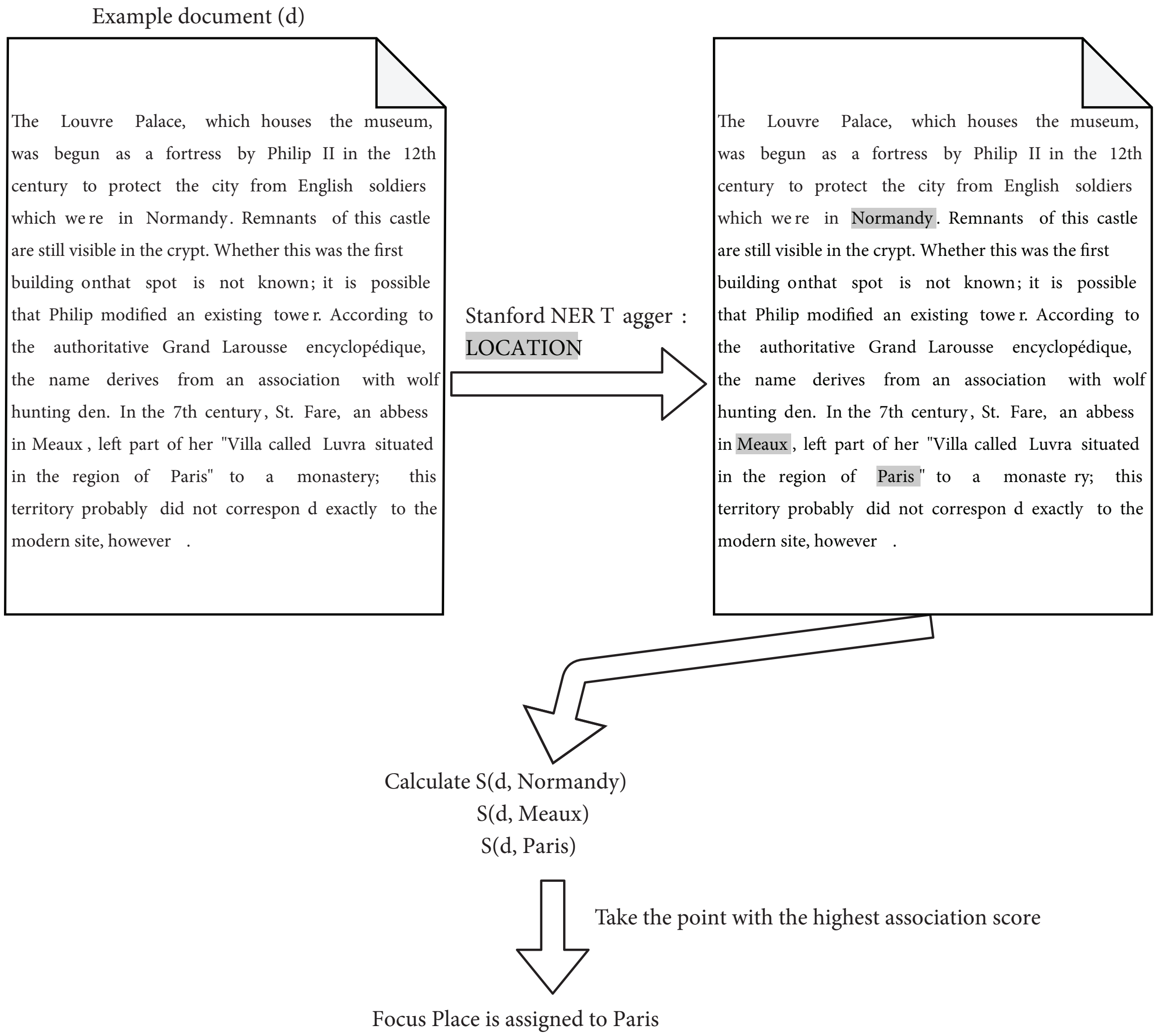
In order to assign a document focus place, we choose a single space point with the highest association score as the estimated focus place. We denote this as $p_{foc}^{ins}(d)$ in Eq. (10):

$$p_{foc}^{ins}(d) = argmax_p S_U(d, p) \tag{10}$$

The procedure of estimating the spatial focus of a given text document is demonstrated in Figure 2.

[11] GeoNames (2006). GeoNames Gazetteer [online] Website http://download.geonames.org/export/dump/ [accessed 20 March 2018].

The example document gives historical information about the Louvre Palace. The Stanford NER Tagger labels Normandy, Meaux, and Paris as locations. In the location estimation calculation using the proposed approach, Normandy and Meaux get almost the same total score. Paris, on the other hand, gets a substantially higher score. This can be explained by the fact that Paris has distinctively higher PMI scores with Louvre, palace, house, museum, castle, villa, etc. in comparison to the corresponding pairwise scores for the other two.



The Louvre Palace, which houses the museum, was begun as a fortress by Philip II in the 12th century to protect the city from English soldiers which we re in Normandy. Remnants of this castle are still visible in the crypt. Whether this was the first building onthat spot is not known; it is possible that Philip modified an existing towe r. According to the authoritative Grand Larousse encyclopédique, the name derives from an association with wolf hunting den. In the 7th century, St. Fare, an abbess in Meaux , left part of her "Villa called Luvra situated in the region of Paris" to a monastery; this territory probably did not correspon d exactly to the modern site, however .

The Louvre Palace, which houses the museum, was begun as a fortress by Philip II in the 12th century to protect the city from English soldiers which we re in Normandy. Remnants of this castle are still visible in the crypt. Whether this was the first building onthat spot is not known; it is possible that Philip modified an existing towe r. According to the authoritative Grand Larousse encyclopédique, the name derives from an association with wolf hunting den. In the 7th century, St. Fare, an abbess in Meaux , left part of her "Villa called Luvra situated in the region of Paris " to a monaste ry; this territory probably did not correspon d exactly to the modern site, however .

Calculate S(d, Normandy)
S(d, Meaux)
S(d, Paris)
Take the point with the highest association score
Focus Place is assigned to Paris

Figure 2. Spatial focus estimation steps for a Wikipedia article.

3.4.2. Dataset groups

In order to test our proposed spatial focus estimation models, we use four different datasets, namely the Wing and Baldridge Wikipedia Test Set, Flickr, Twitter, and Medieval2016. We choose these datasets due to the fact that state-of-the-art approaches are evaluated on these mentioned test sets, and we want to compare our estimation success with those methods.

Wing and Baldridge [36] made available a Wikipedia training and test dataset that consists of 390,574 training and 48,589 test articles. The dataset was taken from the English Wikipedia dump of 4 September 2010 and preprocessed as described in [36]. A tolerable modified version of this dataset was used in [18, 37]. By using this dataset, we can compare the results of our methods to the results reported in [18, 36, 37].

We chose the Medieval2016 PT dataset for the comparison of the performance of our methods to the ones proposed in [17, 27, 38], since [27] made a comparative analysis between their work and [17, 38] using this dataset. Medieval2016 PT is a subset of the YFCC100M dataset [39], consisting of 1,527,398 items.

3.4.3. Results

For the evaluation of the proposed approaches, accuracy (Acc) in various ranges R (Acc@R) and median distance error are used. We count our estimation as correct if the estimated location of a document has been placed within 1, 10, 100, 1000, 10,000, 100,000, or 1,000,000 m of the ground truth location. Estimation distance errors for test items are calculated in kilometers with respect to the distance between the predicted and the actual locations.

The calculation of geodesic distance between the estimated and the real location of an item is based on Karney's algorithm [40]. This algorithm, which relies on the assumption that the shape of the Earth is an oblate spheroid, produces more accurate distances than the methods that assume the shape of Earth is spherical.

State-of-the-art and our results are presented in a combined table for a more comprehensive view of the performance of all methods (Table 4). As can be observed, spatial entropy with PMI outperforms state-of-the-art geoparsing approaches.

Table 4. Geotagging accuracies (percent) of Medieval2016 PT dataset for four ranges and minimal median error of the proposed approach and the state-of-the-art geoparsing approaches.

Method	Acc@0.1 km	Acc@1 km	Acc@10 km	Acc@100 km	Median error
Zhang (optimal)	1.77	13.71	37.04	48.68	131
Zhang (random)	0.68	5.78	17.65	27.87	1148
DBpedia Spotlight (optimal)	1.78	10.94	29.49	37.88	891
DBpedia Spotlight (random)	1.31	8.74	25.22	34.05	1151
Kordopatis-Zilos	7.52	27.40	47.86	56.06	15
Spatial entropy with PMI	9.24	32.78	54.62	63.11	11

We also test the effectiveness of our proposed spatial estimation method by using Flickr and Twitter datasets. In Table 5, accuracy rate and median error of the proposed method for seven different ranges are shown. From the table, it can also be seen that as the range gets wider, such as 1000 km, the accuracy gets higher. Because we are looking at a much wider range, the probability that we accurately predict the location of a document in that range is high. Accuracy values belonging to Flickr along all seven ranges are significantly higher than those of Twitter because the estimation is based mainly on image captions on Flickr data and location information in these captions is rich and mostly correct.

In Table 6, the comparison of success between our spatial estimation approaches and state-of-the-art geoparsing approaches is demonstrated on the Wing and Baldridge Wikipedia Test Set. Roller et al.'s [37] accuracy rates are given in [18]; therefore, we use the same results. For Laere et al.'s [18] accuracy results, we take the most successful ones among the three language models they proposed (Wikipedia, Flickr, and Twitter) in each accuracy range. Our proposed spatial estimation method is shown to be highly competitive with the state-of-the-art on the Wing and Baldridge Wikipedia Test Set as well.

Table 5. Geotagging accuracies (Percent) of Flickr and Twitter dataset for seven ranges and minimal median error.

	Flickr	Twitter
Acc@0.001 km	0.26	0.08
Acc@0.01 km	1.21	0.17
Acc@0.1 km	14.27	0.84
Acc@1 km	40.18	10.85
Acc@10 km	74.03	47.31
Acc@100 km	96.89	75.67
Acc@1000 km	99.04	95.18
Median error	2.35 km	21.48 km

Table 6. Geotagging accuracies (percent) comparison on Wing and Baldrigde Wikipedia test set between the proposed approaches and the state-of-the-art Geoparsing approaches for seven ranges and minimal median error.

	Roller et al.	Laere et al.	Spatial entropy with PMI
Median Error	8.12 km	2.44 km	3.08 km
Acc@0.001 km	0.02%	0.34%	0.49%
Acc@0.01 km	0.02%	0.95%	0.93%
Acc@0.1 km	0.10%	11.52%	10.83%
Acc@1 km	4.17%	37.73%	34.31%
Acc@10 km	53.11%	72.44%	70.22%
Acc@100 km	75.98%	96.47%	96.45%
Acc@1000 km	92.36%	98.84%	98.94%

Additionally, in Table 7, the accuracies of the proposed method (spatial entropy with PMI) are given for all Wikipedia categories that each article is tagged with.

In Table 7, confirming our intuition, the highest accuracy values are obtained for the Wikipedia category of Geography and places, which contain articles with rich geographic contents.

4. Discussion

In temporal focus estimation of documents, the proposed method relies on the local contextual association of terms with time points using PMI. Jatowt et al. [1] used the contextual association of terms with time points based on the Jaccard coefficient. The proposed method's success can be attributed to the use of PMI, because in the case of PMI, a high association means that cooccurrence is bigger than random choice, while the Jaccard coefficient reports high cooccurrence without considering the divergence from randomness. Morbidoni and Cucchiarelli [35] also defined relations among entities and candidate dates and then ranked them to reach the final estimation for focus time. In relating entities with dates, they referred to the occurrence of dates in the full Wikipedia article corresponding to an entity, an entity abstract, RDF temporal triples, etc. When compared to learning word associations through PMI calculations over a large corpus, their way of relating entities to dates is based on presence/counting and does not take into consideration the local context. However, our temporal focus estimation method is a local contextual method and performs better than this global count-based method on diverse datasets of textual documents.

Table 7. Geotagging accuracies (percent) of thirteen categories of Wikipedia articles for seven ranges and minimal median error (spatial entropy with PMI).

Wikipedia categories	Acc@0.001 km	Acc@0.01 km	Acc@0.1 km	Acc@1 km	Acc@10 km	Acc@100 km	Acc@1000 km	Median error
General reference	0.42	0.58	2.13	11.87	60.54	94.23	98.13	6.5 km
Culture and the arts	0.34	0.47	1.95	10.37	57.31	93.43	97.28	6.86 km
Geography and places	0.76	2.34	4.65	20.92	70.54	95.76	98.77	4.03 km
Health and fitness	0.28	0.40	2.05	10.23	58.47	93.23	94.03	9.71 km
History and events	0.42	1.58	3.91	18.58	67.12	95.37	98.46	5.23 km
Human activities	0.63	1.46	2.97	15.73	64.29	94.88	98.34	4.97 km
Mathematics and logic	0.37	0.58	2.76	11.38	57.74	94.38	95.01	7.24 km
Natural and physical sciences	0.41	0.46	2.85	11.29	58.53	94.17	94.44	7.89 km
People and self	0.62	1.75	3.57	15.76	63.32	94.89	97.07	6.92 km
Philosophy and thinking	0.35	0.58	3.11	11.80	60.54	94.23	94.82	7.66 km
Religion and belief systems	0.47	0.63	3.08	12.64	61.89	94.72	96.36	7.18 km
Society and social sciences	0.42	0.60	2.13	11.87	60.58	94.23	95.93	7.36 km
Technology and applied sciences	0.33	0.51	1.83	10.37	56.37	91.77	94.35	8.04 km

Feature selection is important in geocoding. The baseline work [17] in this area depends on the combination of gazetteer features, surface similarity with respect to morphology, twitter context, and metadata features in classifiers. Recent language modeling-based approaches have given better results. Among these, Kordopatis-Zilos et al. [27] created a rectangular grid of cells and generated term-cell probabilities. The term occurrence probabilities are calculated from processing a geotagged corpus. It can be concluded that approaches that consider term occurrences in the context of a location term prove more useful in estimating a location for a text. The proposed approach is in a similar direction. It uses PMI to associate terms with place points. Laere et al.'s work [18] is another probabilistic language modeling approach that exploits correlations between the occurrence of terms and locations. Distinctively, their model is trained on Flickr and Twitter in addition to Wikipedia. Their superior performance on Wing and Baldridge Wikipedia Test Set indicates that social media sites are useful resources of geographical information.

As a final remark, in both temporal and spatial estimation, local context-based considerations introduce performance boosts. Thus, neural network methods that learn representations from local contexts look promising in this area. A transition from feature selection to feature representation learning might be expected.

5. Conclusion and future work

Time and place, mattering greatly in our lives, appear in text documents frequently. Being able to automatically categorize documents by their spatiotemporal focus is important because it will improve spatiotemporal information retrieval and help to better analyze and understand documents.

In this paper, we propose temporal entropy calculation using PMI to estimate the focus time of documents. Unlike the existing work, we propose to use PMI rather than the Jaccard coefficient to calculate word association scores, since PMI further normalizes the scores by using probability values, whereas the Jaccard coefficient uses only raw frequencies.

As another contribution of this paper, we adapt the proposed approach for estimating temporal focus to spatial focus estimation. Despite the inherent differences between time and location estimation, as can be observed from the geotagging accuracy results, the proposed spatial focus estimation approach is highly competitive.

Moreover, our methods for spatiotemporal focus estimation are also able to work when documents do not have explicit time expressions or location entities as we learn word-time point and word-place point associations from PMI calculations over a Wikipedia corpus.

In the case of spatial estimation, in addition to Wikipedia, PMI scores can be computed over Flickr and Twitter because these social media sites are rich resources of geographical information.

As word embeddings learn local contextual information and they are shown to provide substantial performance improvement in downstream natural language processing tasks, we plan to use word embeddings for the problem of spatiotemporal estimation.

Acknowledgment

We would like to thank Assoc. Prof. Tuğkan Tuğlular for his useful comments on versions of the manuscript.

References

[1] Jatowt A, Au Yeung C, Tanaka K. Generic method for detecting focus time of documents. Information Processing & Management 2015; 51 (6): 851-868.

[2] Strötgen J, Armiti A, Van Canh T, Zell J, Gertz M. Time for more languages: Temporal tagging of Arabic, Italian, Spanish, and Vietnamese. ACM Transactions on Asian Language Information Processing 2014; 13 (1): 1.

[3] De Jong F, Rode H, Hiemstra D. Temporal language models for the disclosure of historical text. In: Humanities, Computers and Cultural Heritage: Proceedings of the 16th International Conference of the Association for History and Computing; Amsterdam, the Netherlands; 2005. pp. 161-168.

[4] Kanhabua N, Nørvåg K. Improving temporal language models for determining time of non-timestamped documents. In: International Conference on Theory and Practice of Digital Libraries; Aarhus, Denmark; 2008. pp. 358-370.

[5] Kanhabua N, Nørvåg K. Using temporal language models for document dating. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases; Bled, Slovenia; 2009. pp. 738-741.

[6] Kumar A, Lease M, Baldridge J. Supervised language modeling for temporal resolution of texts. In: 20th ACM International Conference on Information and Knowledge Management; Glasgow, UK; 2011. pp. 2069-2072.

[7] Kumar A, Baldridge J, Lease M, Ghosh J. Dating texts without explicit temporal cues. arXiv preprint, arXiv:1211.2290, 2012.

[8] Dalli A. Temporal classification of text and automatic document dating. In: Human Language Technology Conference of the NAACL, Companion Volume: Short Papers; New York City, USA; 2006. pp. 29-32.

[9] Chambers N. Labeling documents with timestamps: learning from their time expressions. In: 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1; Jeju Island, Korea; 2012. pp. 98-106.

[10] Niculae V, Zampieri M, Dinu L, Ciobanu AM. Temporal text ranking and automatic dating of texts. In: 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers; Gothenburg, Sweden; 2014. pp. 17-21.

[11] Garcia-Fernandez A, Ligozat AL, Dinarelli M, Bernhard D. When was it written? Automatically determining publication dates. In: International Symposium on String Processing and Information Retrieval; Pisa, Italy; 2011. pp. 221-236.

[12] Kotsakos D, Lappas T, Kotzias D, Gunopulos D, Kanhabua N et al. A burstiness-aware approach for document dating. In: 37th International ACM SIGIR Conference on Research & Development in Information Retrieval; Gold Coast, Australia; 2014. pp. 1003-1006.

[13] Melo F, Martins B. Automated geocoding of textual documents: a survey of current approaches. Transactions in GIS 2017; 21 (1): 3-38.

[14] Woodruff AG, Flaunt C. Automated Geographic Indexing of Text Documents (Sequoia 2000 Technical Report 94/41). Berkeley, CA: University of California, EECS, 1994.

[15] Martins B, Manguinhas H, Borbinha J. Extracting and exploring the geo-temporal semantics of textual resources. In: IEEE International Conference on Semantic Computing; Washington, DC, USA; 2008. pp. 1-9.

[16] Han B, Cook P, Baldwin T. Geolocation prediction in social media data by finding location indicative words. In: 24th International Conference on Computational Linguistics; Mumbai, India; 2012. pp. 1045-1062.

[17] Zhang W, Gelernter J. Geocoding location expressions in Twitter messages: a preference learning method. Journal of Spatial Information Science 2014; 2014 (9): 37-70.

[18] Van Laere O, Schockaert S, Tanasescu V, Dhoedt B, Jones CB. Georeferencing Wikipedia documents using data from social media sources. ACM Transactions on Information Systems 2014; 32 (3): 12.

[19] Wing B, Baldridge J. Hierarchical discriminative classification for text-based geolocation. In: Conference on Empirical Methods in Natural Language Processing; Doha, Qatar; 2014. pp. 336-348.

[20] Priedhorsky R, Culotta A, Del Valle SY. Inferring the origin locations of tweets with quantitative confidence. In: 17th ACM Conference on Computer Supported Cooperative Work & Social Computing; New York, NY, USA; 2014. pp. 1523-1536.

[21] Li G, Hu J, Feng J, Tan KL. Effective location identification from microblogs. In: 30th International Conference on Data Engineering; Chicago, IL, USA; 2014. pp. 880-891.

[22] Van Laere O, Quinn J, Schockaert S, Dhoedt B. Spatially aware term selection for geotagging. IEEE Transactions on Knowledge and Data Engineering 2013; 26 (1): 221-234.

[23] Rahimi A, Vu D, Cohn T, Baldwin T. Exploiting text and network context for geolocation of social media users. arXiv preprint, arXiv:1506.04803, 2015.

[24] Hulden M, Silfverberg M, Francom J. Kernel density estimation for text-based geolocation. In: 29th AAAI Conference on Artificial Intelligence; Austin, TX, USA; 2015. pp. 145-150.

[25] Brunsting S, De Sterck H, Dolman R, Van Sprundel T. GeoTextTagger: High-Precision Location Tagging of Textual Documents using a Natural Language Processing Approach. arXiv preprint, arXiv:1601.05893, 2016.

[26] Rodrigues E, Assunção R, Pappa GL, Renno D, Meira W Jr. Exploring multiple evidence to infer users' location in Twitter. Neurocomputing 2016; 171: 30-38.

[27] Kordopatis-Zilos G, Papadopoulos S, Kompatsiaris I. Geotagging text content with language models and feature mining. Proceedings of the IEEE 2017; 105 (10): 1971-1986.

[28] Strötgen J, Gertz M. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In: 5th International Workshop on Semantic Evaluation; Uppsala, Sweden; 2010. pp. 321-324.

[29] Setzer A. Temporal information in newswire articles: an annotation scheme and corpus study. PhD, University of Sheffield, Sheffield, UK, 2001.

[30] Ferro L, Mani I, Sundheim B, Wilson G. TIDES Temporal Annotation Guidelines Version 1.0.2. McLean, VA, USA: MITRE Corporation, 2001.

[31] Grishman R, Sundheim B. Message understanding conference-6: A brief history. In: 16th Conference on Computational Linguistics - Volume 1; Copenhagen, Denmark; 1996. pp. 466-471.

[32] Jaccard P. The distribution of the flora in the alpine zone. 1. New Phytologist 1912; 11 (2): 37-50.

[33] Church KW, Hanks P. Word association norms, mutual information, and lexicography. Computational Linguistics 1990; 16 (1): 22-29.

[34] Mazur P, Dale R. Wikiwars: A new corpus for research on temporal expressions. In: Conference on Empirical Methods in Natural Language Processing; Cambridge, MA, USA; 2010. pp. 913-922.

[35] Morbidoni C, Cucchiarelli A. A bag-of-entities approach to document focus time estimation. In: 3rd International Workshop on Knowledge Discovery on the WEB; Cagliari, Italy; 2017.

[36] Wing BP, Baldridge J. Simple supervised document geolocation with geodesic grids. In: 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1; Portland, OR, USA; 2011. pp. 955-964.

[37] Roller S, Speriosu M, Rallapalli S, Wing B, Baldridge J. Supervised text-based geolocation using language models on an adaptive grid. In: 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning; Jeju Island, Korea; 2012. pp. 1500-1510.

[38] Daiber J, Jakob M, Hokamp C, Mendes PN. Improving efficiency and accuracy in multilingual entity extraction. In: 9th International Conference on Semantic Systems; Graz, Austria; 2013. pp. 121-124.

[39] Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K et al. YFCC100M: The new data in multimedia research. arXiv preprint, arXiv:1503.01817, 2015.

[40] Karney CF. Algorithms for geodesics. Journal of Geodesy 2013; 87 (1): 43-55.