Supervised Language Modeling for Temporal Resolution of Texts

Abhimanu Kumar
Dept. of Computer Science
University of Texas at Austin
abhimanu@cs.utexas.edu

Matthew Lease
School of Information
University of Texas at Austin
ml@ischool.utexas.edu

Jason Baldridge
Department of Linguistics
University of Texas at Austin
jbaldrid@mail.utexas.edu

ABSTRACT

We investigate *temporal resolution* of documents, such as determining the date of publication of a story based on its text. We describe and evaluate a model that build histograms encoding the probability of different temporal periods for a document. We construct histograms based on the Kullback-Leibler Divergence between the language model for a test document and supervised language models for each interval. Initial results indicate this language modeling approach is effective for predicting the dates of publication of short stories, which contain few explicit mentions of years.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

time, temporal information, text mining, document dating

1. INTRODUCTION

There is a long tradition of work on temporal analysis of texts. In computational linguistics, the primary focus has been on the fine-grained level of temporal ordering of individual events necessary for deep natural language understanding [1, 24]. In Information Retrieval (IR), research has investigated time-sensitivity document ranking [9, 17], time-based organization and presentation of search results [2], how queries and documents change over time [16], etc.

This paper describes temporal analysis of documents using methods rooted in both computational linguistics and IR. While accurate extraction and resolution of explicit mentions of time (absolute or relative) is clearly important [2], our primary focus here is the challenge of learning implicit temporal distributions over natural language use at-large. As a concrete example, consider the word *wireless*. Introduced in the early 20th century to refer to radios, it fell into disuse and then its meaning shifted later that century to describe any form of communication without cables (e.g. internet access). As such, the word *wireless* embodies implicit time cues, a notion we might generalize by inferring its complete temporal distribution as well as that of all other words. By harnessing many such implicit cues in combination across a document, we might further infer a unique temporal distribution for the overall document.

As in prior document dating studies, we partition the timeline (and document collection) to infer an unique language model (LM) underlying each time period [10, 14]. While prior work considered texts from the past 10-20 years, our work is more historically-oriented, predicting publication dates for historical works of fiction.

After estimating a similar LM underlying each document [20], we measure similarity between each document's LM and each time period's LM, yielding a distribution over the timeline for each document [10, 14]. In addition to document dating applications, this distribution has potential to inform work in computational humanities, specifically scholars' understandings of how a work was influenced by or reflects different time periods.

We predict publication dates of short stories from the Gutenberg project[1] published between 1798 to 2008. Gutenberg stories are labeled by publication year. Our inference task is then to predict the publication year given the story's text. These short works of fiction typically use relatively few explicit temporal expressions or mentions of real-life named entities. We refer to the stories as *documents*, with Gutenberg defining a document *collection* c consisting of N documents: $c = d_{1:N}$. Our best model achieves a median error of 23 years from the true publication dates.

2. RELATED WORK

Annotation and corpora. Recent years have brought increased interest in creating new, richly annotated corpora for training and evaluating time-sensitive models. TimeBank [21] and Wikiwars [19] are great exemplars of such work. They have been used for tasks like modeling event structure (e.g. work of [7] on TimeBank).

Document dating. Prior work on LM-based document dating [10, 14] partitioned the timeline (and document collection) to infer a unique LM underlying each time period. A LM underlying each document [20] was also estimated and used to measure similarity vs. each time period's LM, yielding a distribution over the timeline for each document. However, while prior work focused on the past 10-20 years, our work is more historically-oriented, modeling the timeline from the present day back to the 18th century.

Foundational work by de Jong et al. [10] considered Dutch newspaper articles from 1999-2005 and compared language models using the normalised log-likelihood ratio measure (NLLR), a variant of KL-divergence. Linear and Dirichlet smoothing were ap-

[1] http://www.gutenberg.org

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

plied, apparently to the partition LMs but not the document LMs. They also distinguish between output granularity of time (to be predicted) and the granularity of time modeled. Kanhabua et al. [14] extended de Jong et al.'s model with notions of temporal entropy, use of search term trends from Google Zeitgeist, and semantic preprocessing. Temporal entropy weights terms differently based on how well a term distinguishes between time partitions and how important a term is to a specific partition. Semantic techniques included part-of-speech tagging, collocation extraction, word sense disambiguation, and concept extraction. They created a time-labeled document collection by downloading web pages from the Internet Archive spanning an eight year period. In follow-on work inferring temporal properties of queries [15], they used the New York Times annotated corpus, with articles spanning 1987-2007.

IR Applications. IR research has investigated time-sensitivity query interpretation and document ranking [17, 9], time-based organization and presentation of search results [2], how queries and documents change over time [16], etc. One of the first LM-based temporal approaches by Li and Croft [17] used explicit document dates to estimate a more informative document prior. More recent work by Dakka et al. [9] automatically identify important time intervals likely to be of interest for a query and similarly integrate knowledge of document publication date into the ranking function. The most relevant work to ours is that by Alonso et al. [2], who provide valuable background on motivation, overview and discussion of temporal analysis in IR. Using explicit temporal metadata and expressions, they create *temporal document profiles* to cluster documents and create timelines for exploring search results.

Time-sensitive topic modeling. There has been a variety of work on time based topic-analysis in texts in recent years, such as Dynamic Topic Models [6]. Subsequent work [5] proposes probabilistic time series models to analyze the time evolution of topics in a large document collection. They take a sequential collection of documents of a particular area e.g. news articles and determine how topics evolve over time - topics appearing and disappearing, new topics emerging and older ones fading away. [25] provide a model to evaluate variations in the occurrence of topics in large corpora over a period of time. There have been other interesting contributions such as work by [18] which studies the history of ideas in a research field using topic models, by [8] which provides the temporal analysis of blogs and by [27] which gives models for mining cluster evaluation from time varying text corpora.

Geolocation. Temporal resolution can be seen as a natural pairing with geolocation: both are ways of connecting texts to simple, but tremendously intuitive and useful, models of aspects of the real world. There has been a long-standing interest in finding ways to connect documents to specific places on the earth, especially for geographic information retrieval [12, 3]. Of particular relevance to our paper is Wing and Baldridge's LM based method of measuring similarity of documents with language models for discrete geodesic cells on the earth's surface [26].

Authorship attribution. A final relevant work is the LM-based authorship attribution work by Zhao et al. [28]. They similarly partition the corpus by author, build partition-specific LMs, and infer authorship based on model similarity computed with KL-divergence and Dirichlet smoothing. They also consider English literature from the Gutenberg Project. Unlike us, they train directly on this corpus instead of applying LMs from another domain.

3. MODELING AND ESTIMATION

Following aforementioned prior work [2, 10, 14], we quantize continuous time into discrete units. Our terminology and formalization most closely follow that of Alonso et al. [2]. The smallest temporal granularity we consider in this work is a single year, though the methods we describe can in principle be used with units of finer granularity such as days, weeks, months, etc.

A *chronon* is an atomic interval x upon which a discrete timeline is constructed [2]. In this paper, a chronon consists of δ years, where δ is a tunable parameter. Given δ, the timeline T_δ is decomposed into a sequence of n contiguous, non-overlapping chronons $\mathbf{x} = x_{1:n}$, where $n = \frac{\Delta}{\delta}$.

We model the affinity between each chronon x and a document d by estimating the discrete distribution $P(x|d)$, a document-specific normalized histogram over the timeline of chronons $x_{1:n}$. In the next section, we use $P(x|d)$ to infer affinity between d and different chronons as well as longer granules. We describe an LM-based approach for inferring affinity between chronon x and document d as a function of model divergence between latent unigram distributions $P(w|d)$ and $P(w|x)$ (similar to [10, 14]).

We model the likelihood of each chronon x for a given document d. By forming a unique "pseudo-document" d^x associated with each chronon x, we estimate Θ^x from d^x and estimate $P(x|d)$ by comparing the similarity of Θ^d and Θ^x [10, 14]. Whereas prior work measured similarity via NLLR (§2), we compute the unnormalized likelihood of some x given d via standard (inverse) KL-divergence $\mathcal{D}(\Theta^d||\Theta^x)^{-1}$ and normalize in straight-forward fashion over all chronons $x_{1:n}$:

$$P_{wa}(x|d) = \frac{\mathcal{D}(\Theta^d||\Theta^x)^{-1}}{\sum_x \mathcal{D}(\Theta^d||\Theta^x)^{-1}} \quad (1)$$

We generate the "pseudo-document" d^x for each chronon x by including all training documents whose labeled span overlaps x.

As in prior work [10, 14, 28], we adopt Dirichlet smoothing to regularlize partition LMs. However, rather than adopt a single fixed prior for all documents and chronons, we instead define document-chronon specific priors. Let $|V_{d^x} \cup V_d|$ denote the document-chronon specific vocabulary for some collection document d_i and pseudo-document d^x. For each document-chronon pair, we define the prior to be a uniform distribution over this specific vocabulary: $\frac{1}{|V_{d^x} \cup V_{d_i}|}$. For any collection document or pseudo-document d, we perform Dirichlet smoothing of form:

$$\hat{\theta_w^d} = \eta \frac{f_w^d}{|d|} + (1-\eta)\frac{1}{|V_{d^x} \cup V_d|}\ ,\ \eta = \frac{|d|}{|d|+\mu} \quad (2)$$

Rather than specify the hyper-parameter μ directly, however, we introduce another hyper-parameter λ, where $\mu = \frac{\lambda}{|V_{d^x} \cup V_{d_i}|}$.

The intuition for this smoothing is that mass is provided only for the words that are present in the collection document or the chronon's pseudo-document, ignoring other words. Thus, the divergence calculation is done only with respect to either the document or the chronon we have evidence for. There are many chronons for which we have little textual evidence; if these are smoothed with respect to all words in the collection, then those terms dominate the divergence calculation. When a short document is evaluated against a low-evidence chronon, smoothing over all words leads to many terms (few of which actually occur in the document or the chronon) having similar probabilities, leading to low divergence.

To infer a representative chronon x^d for each document, **MAXCHRONON** simply selects the most-likely chronon under $P(x|d)$:

$$\mathbf{x_{mc}^d} = \arg\max_{x \in \mathbf{x}} P(x|d) \quad (3)$$

We use MAXCHRONON to predict a chronon for the document and then select the first year in that chronon as the predicted year.

4. EVALUATION

Data. We use 678 Gutenberg short stories, split into development and test sets of 333 and 345 each, respectively. Strictly speaking, this split is not actually necessary for our reported experiments since we tune out-of-domain on a separate corpus (see below). Nonetheless, we also plan to tune in-domain in our future work, so we have gone ahead and used the division now for direct comparison to later results. The average, minimum and maximum word count of these stories are 14k, 11k and 100k respectively. All numeric tokens and standard stopwords are removed.

Wikipedia Biographies. We tune parameters using an external corpus of biographies from the September 4, 2010 dump of English Wikipedia[2] (our use of Wikipedia here is motivated by other preliminary work on it to be further developed in our future work). We consider individuals whose complete lives occurred within the year range 3800 B.C. to 2010 A.D. We extract the lifetime of each individual via each article's *Infobox* `birth_date` and `death_date` fields. We exclude biographies which do not specify both fields or which fall outside the year range considered. Given the lifetime of each individual, we take the midpoint of individual's lifetime as the gold standard year to match for the biography. We note that as is often typical of Wikipedia coverage, the distribution of biographies is quite skewed toward recent times.

As examples of the kinds of distributions we obtain for Wikipedia biographies, Figure 1 shows graphs of $P(x|d)$ for (a) Plato and (b) Abraham Lincoln. Recall that these are based on no explicit temporal expressions. For Plato, there is a clear spike around the time he was alive, along with another rising bump toward the current day, reflecting modern interest in him. For Lincoln, there is a single peak at 1835—very close to the 1837 midpoint of his life.

Parameters and Estimation. We set WORDAFFINITY's smoothing parameter $\lambda = 1$ without any tuning (left for future work). The chronon size δ was tuned through experimentation to optimize performance on the Wikipedia biography collection. Two salient points of note are: (a) the task will be more challenging since we have a tune/test corpus mismatch between Wikipedia vs. Gutenberg, and (b) a particular challenge of this is differences in vocabulary selection between recently written Wikipedia articles and historical fiction in the Gutenberg stories.

As in prior work [10, 14], we smooth chronon pseudo-document language models but not document models. While smoothing both may potentially help, smoothing the former is strictly necessary to prevent division by zero in the KL-divergence calculation.

Year Prediction. The gold standard to match for each document is a labeled year (publication year for Gutenberg, lifetime midpoint for Wikipedia). Given document d, we predict the year to be the first year from x^d_{mc}, the most-likely chronon for d over the timeline.

When predicting a single year for a document, a natural error measure between the predicted year $\hat{y}$ and the actual year y^* is the difference $|\hat{y} - y^*|$. We compute this difference for each document, then compute and report the mean and median of differences across documents. Similar distance error measures have also been used with document geolocation [13, 26].

Baseline. As a simple baseline, we consider a fixed prediction of the year 1903 for all stories: the midpoint of publication date range (1798-2008) in the collection. This assumes that one knows a rough range of possible publication dates, which might be reasonable in many cases and provide a useful starting point for comparison.

Development Set Tuning and Results. Using the Wikipedia biographies, we tune $\delta \in \{1, 2, \ldots, 100\}$ and examine mean error in predicted year. Error roughly varies between 85-155, with $\delta = 40$ seen to yield optimal mean year prediction error (see Figure 2). Next, we use $\delta = 40$ on the Gutenberg collection. In comparison to the 1903 baseline, mean prediction error is reduced from 36 to 28 years, with median prediction error cut from 50 to 32 years.

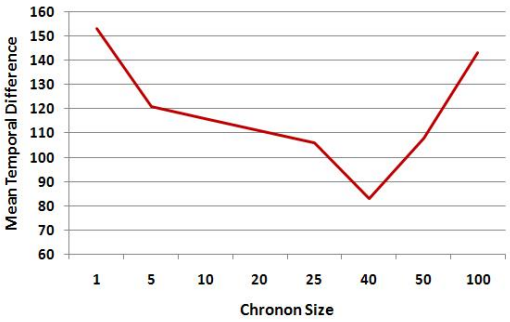

Figure 2: Chronon size δ vs. resultant mean error of year predictions for lifespan midpoints in Wikipedia biographies.

Test Set Results. While results show only a modest reduction of mean error in year predictions for the model vs. the 1903 baseline (37 to 34), median error is more than halved, dropping from 50 to 23. This provides a promising validation of our initial work that words which are not temporal expressions still convey important temporal information that can be exploited for historical document dating, even when (a) the model was trained on a different domain and (b) there is only a range of 210 years under consideration and the baseline represents the exact midpoint.

5. CONCLUSION

We have shown that it is possible to perform accurate temporal resolution of texts by combining evidence from both explicit temporal expressions and the implicit temporal properties of general words. We create a partitioned timeline with language models in each partition; these are used to determine a probability over the timeline for new test documents. The models are effective enough that they can predict publication dates on out-of-domain stories from a 210 year period to within 32 years on average.

There are a number of ways to improve the present approach. For example, we might smooth LMs for the chronons themselves such that we infer similar temporal distributions proportional to proximity between chronons. Similar intuition has appeared in LM-based IR work in various forms: similar queries should have similar LMs, similar documents should have similar LMs [23], and similar documents should receive similar scores when compared to to some other fixed model [11].

Another, obvious idea is to consider n-grams rather than unigrams: e.g., *civil war* has different temporal properties from *civil* and *war* on their own. Ultimately, one of our interests in this general line of inquiry is to create models of word meaning that are *grounded*. In most computational models of word meaning, a word is characterized in terms of other words—a circular (though useful) model. However, there is recent interest in models that connect words to properties of the real world, such as conceptual representations [4] and models of space [22] and time. A relevant task for this idea is word sense disambiguation. For example, the word *apple* could be a reference to the fruit (general over time), the record company formed by the Beatles (1960-1970s, primarily), or Apple Inc. (1976-present). Identifying the temporal properties of the

[2] `http://download.wikimedia.org/enwiki/20100904/enwiki-20100904-pages-articles.xml.bz2`

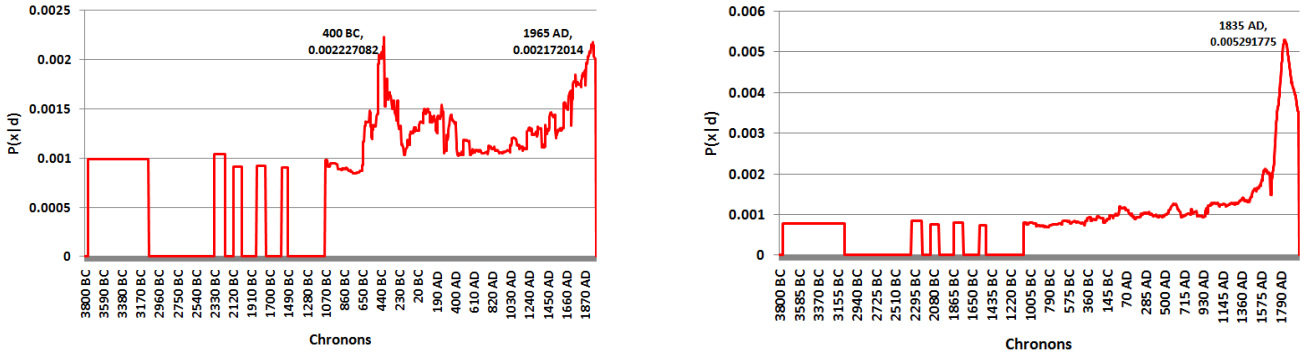

Figure 1: Example $P(x|d)$ distributions for the biographies of (a) Plato (428-348 B.C.) and (b) Abraham Lincoln (1809-1965).

text can be part of disambiguating such terms and help us keep a computer or record company from falling on Isaac Newton's head.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback and suggestions. This work was partially supported by a John P. Commons Fellowship for Matthew Lease and by a grant from the Morris Memorial Trust Fund of the New York Community Trust.

6. REFERENCES

[1] J. F. Allen. Maintaining knowledge about temporal intervals. In *Communications of the ACM, Volume 26 Issue 11*, 1983.

[2] O. Alonso, M. Gertz, and R. Baeza-Yates. Clustering and exploring search results using timeline constructions. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 97–106, 2009.

[3] G. Andogah. *Geographically Constrained Information Retrieval*. PhD thesis, University of Groningen, Netherlands, May 2010.

[4] M. Baroni, B. Murphy, E. Barbu, and M. Poesio. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222–254, 2010.

[5] D. Blei, C. Wang, and D. Heckerman. Continuous time dynamic topic models. In *Intl. Conference on Machine Learning (ICML)*, 2008.

[6] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *23rd Intl. Conference on Machine Learning (ICML)*, 2006.

[7] N. Chambers and D. Jurafsky. Jointly combining implicit constraints improves temporal ordering. In *EMNLP*, 2008.

[8] Y. Chi, S. Zhu, X. Song, J. Tatemura, and B. L. Tseng. Structural and temporal analysis of the blogosphere through community factorization. In *Proc. of 13th SIGKDD*, 2007.

[9] W. Dakka, L. Gravano, and P. Ipeirotis. Answering general time-sensitive queries. *Knowledge and Data Engineering, IEEE Transactions on*, (99), 2010. Pre-print.

[10] F. de de Jong, H. Rode, and D. Hiemstra. Temporal language models for the disclosure of historical text. In *XVIth International Conference of the Association for History and Computing (AHC)*, pages 161–168, 2005.

[11] F. Diaz. Regularizing query-based retrieval scores. *Information Retrieval*, 10(6):531–562, 2007.

[12] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *Proc. of the 26th Intl. Conference on Very Large Data Bases (VLDB)*, pages 545–556, 2000.

[13] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, 2010.

[14] N. Kanhabua and K. Nørvåg. Improving temporal language models for determining time of non-timestamped documents. In *12th European conf. Research and Advanced Technology for Digital Libraries (ECDL)*, pages 358–370, 2008.

[15] N. Kanhabua and K. Nørvåg. Determining time of queries for re-ranking search results. In *14th European conference on Research and Advanced Technology for Digital Libraries (ECDL)*, pages 261–272, 2010.

[16] A. Kulkarni, J. Teevan, K. Svore, and S. Dumais. Understanding temporal query dynamics. In *Proceedings of the 4th ACM International conference on Web Search and Data Mining (WSDM)*, pages 167–176, 2011.

[17] X. Li and W. Croft. Time-based language models. In *12th International Conference on Information and Knowledge Management (CIKM)*, pages 469–475, 2003.

[18] C. D. Manning, D. Hall, and D. Jurafsky. Studying the history of ideas using topic models. In *ACL*, 2008.

[19] P. Mazur and R. Dale. Wikiwars: A new corpus for research on temporal expressions. In *EMNLP*, Massachusetts, 2010.

[20] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281, 1998.

[21] J. Pustejovsky, P. Hanks, R. Sauri, A. See, D. Day, L. Ferro, R. Gaizauskas, M. Lazo, A. Setzer, and B. Sundheim. The TimeBank corpus. *Corpus Linguistics*, pages 647–656, 2003.

[22] M. Speriosu, T. Brown, T. Moon, J. Baldridge, and K. Erk. Connecting language and geography with region-topic models. In *1st Workshop on Computational Models of Spatial Language Interpretation*, 2010.

[23] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pages 407–414, 2006.

[24] M. B. Vilain and B. Beranek. A system for reasoning about time. In *Proceedings of the Second National Conference on Artificial Intelligence (AAAI)*, Pittsburgh, 1982.

[25] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Knowledge Discovery and Data-mining (KDD)*, 2006.

[26] B. Wing and J. Baldridge. Simple supervised document geolocation with geodesic grids. In *ACL-HLT*, 2011.

[27] J. Zhang, Y. Song, and C. Zhang. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *Proceedings of the 16th ACM SIGKDD*, 2010.

[28] Y. Zhao, J. Zobel, and P. Vines. Using relative entropy for authorship attribution. *Third Asia Information Retrieval Symposium (AIRS)*, pages 92–105, 2006.