Mining temporal footprints from Wikipedia

Michele Filannino
School of Computer Science
The University of Manchester
Manchester, M13 9PL, UK
filannim@cs.man.ac.uk

Goran Nenadic
School of Computer Science
The University of Manchester
Manchester, M13 9PL, UK
g.nenadic@manchester.ac.uk

Abstract

Discovery of temporal information is key for organising knowledge and therefore the task of extracting and representing temporal information from texts has received an increasing interest. In this paper we focus on the discovery of temporal footprints from encyclopaedic descriptions. Temporal footprints are time-line periods that are associated to the existence of specific concepts. Our approach relies on the extraction of date mentions and prediction of lower and upper boundaries that define temporal footprints. We report on several experiments on persons' pages from Wikipedia in order to illustrate the feasibility of the proposed methods.

1 Introduction

Temporal information, like dates, durations, time stamps etc., is crucial for organising both structured and unstructured data. Recent developments in the natural language community show an increased interest in systems that can extract temporal information from text and associate it to other concepts and events. The main aim is to detect and represent the temporal flow of events narrated in a text. For example, the TempEval challenge series (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013) provided a number of tasks that have resulted in several temporal information extraction systems that can reliably extract complex temporal expressions from various document types (UzZaman and Allen, 2010; Llorens et al., 2010; Bethard, 2013; Filannino et al., 2013).

In this paper we investigate the extraction of *temporal footprints* (Kant et al., 1998): continuous periods on the time-line that temporally define a concept's existence. For example, the temporal footprint of people lies between their birth and death, whereas temporal footprint of a business company is a period between its constitution and closing or acquisition (see Figure 1 for examples). Such information would be useful in supporting several knowledge extraction and discovery tasks. A question answering system, for example, could spot temporally implausible questions (e.g. *What computer did Galileo Galilei use for his calculations?* or *Where did Blaise Pascal meet Leonardo Da Vinci?*), or re-rank candidate answers with respect to their temporal plausibility (e.g. *British politicians during the Age of Enlightenment*). Similarly, temporal footprints can be used to identify inconsistencies in knowledge bases.

Temporal footprints are in some cases easily accessible by querying Linked Data resources (e.g. DBPedia, YAGO or Freebase) (Rula et al., 2014), large collections of data (Talukdar et al., 2012) or by directly analysing Wikipedia info-boxes (Nguyen et al., 2007; Etzioni et al., 2008; Wu et al., 2008; Ji and Grishman, 2011; Kuzey and Weikum, 2012). However, the research question we want to address in this paper is whether it is possible to automatically approximate the temporal footprint of a concept only by analysing its encyclopaedic description rather than using such conveniently structured information.

This paper is organised as follows: Section 2 describes our approach and four different strategies to predict temporal footprints. Section 3 provides information about how we collected the data for the experiments, and Section 4 presents and illustrates the results.

Proceedings of the AHA! Workshop on Information Discovery in Text, pages 7–13,
Dublin, Ireland, August 23 2014.

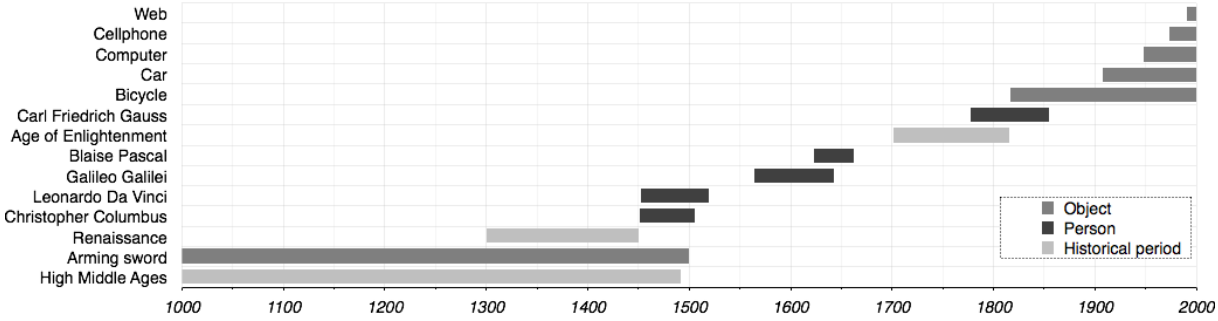


Figure 1: Examples of temporal footprints of objects, people and historical periods.

2 Methodology

In order to identify a temporal footprint for a given entity, we propose to predict its lower and upper bound using temporal expressions appearing in the associated text. The approach has three steps: (1) extracting mentions of temporal expressions, (2) filtering outliers from the obtained probability mass function of these mentions, and (3) fitting a normal distribution to this function. This process is controlled by three parameters we introduce and describe below. We restrict temporal footprints to the granularity of years.

2.1 Temporal expression extraction (TEE)

For each concept we extract all the dates from its associated textual content (e.g. a Wikipedia page). There are numerous ways to extract mentions of dates, but we use (a) regular expressions that search for mentions of full years (e.g. sequence of four digits that start with '1' or '2' (e.g. *1990*, *1067* or *2014*) — we refer to this as TEE RegEx; (b) a more sophisticated temporal expression extraction system, which can also extract implicit date references, such as *"a year after"* or *"in the same period"*, along with the explicit ones and, for this reason, would presumably be able to extract more dates. As temporal expression extraction system we used HeidelTime (Strötgen et al., 2013), the top-ranked in TempEval-3 challenge (UzZaman et al., 2013). We refer to this approach as TEE Heidel.

2.2 Filtering (Flt)

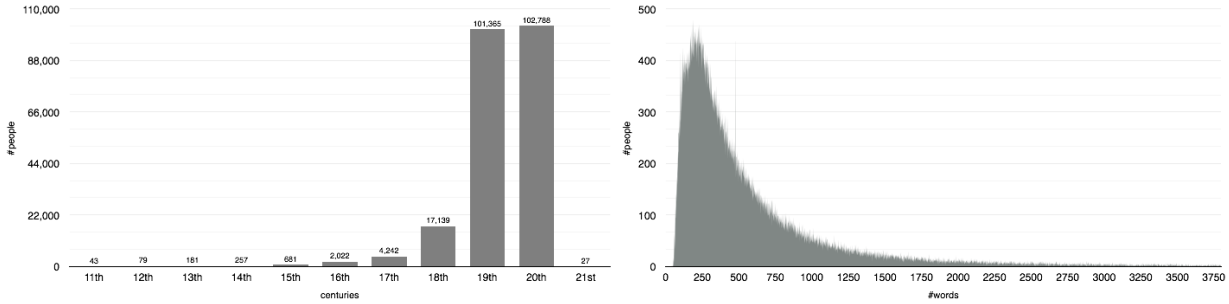
We assume that the list of all extracted years gives a probability mass function. We first filter outliers out from it using the Median Absolute Deviation (Hampel, 1974; Leys et al., 2013) with a parameter (γ) that controls the size of the acceptance region for the outlier filter. This parameter is particularly important to filter out present and future references, invariably present in encyclopaedic descriptions. For example, in the sentence "Volta also studied what we now call electrical capacitance", the word *now* would be resolved to '2014' by temporal expression extraction systems, but it should be discarded as an outlier when discovering of Volta's temporal footprint.

2.3 Fitting normal distribution (FND)

A normal distribution is then fitted on the filtered probability mass function. Lower and upper bounds for a temporal footprint are predicted according to two supplementary parameters, α and β. More specifically, the α parameter controls the width of the normal distribution by resizing the width of the Gaussian bell. The β parameter controls the displacement (shift) of the normal distribution. For example, in the case of Wikipedia pages about people, typically this parameter has a negative value (e.g. -5 or -10 years) since the early years of life are rarely mentioned in an encyclopaedic description. We compute the upper and lower bounds of a temporal footprint using the formula $(\mu + \beta) \pm \alpha\sigma$.

We experimented with the following settings:

(a) The *TEE RegEx* strategy consists of extracting all possible dates by using the regular expression previously mentioned and by assigning to the lower and upper bound the earliest and the latest extracted year respectively.



(a) Distribution of Wikipedia pages per century.

(b) Distribution of Wikipedia pages per length (in words).

Figure 2: Exploratory statistics about the test set extracted from DBpedia.

(b) In the *TEE RegEx + Flt* approach, we first discard outliers from the extracted dates and then the earliest and latest dates are used for lower and upper bounds.

(c) For the *TEE RegEx + Flt + FND* strategy, we use the regular expression-based extraction method and then apply filtering and Gaussian fitting.

(d) Finally, for the *TEE Heidel + Flt + FND* setting, we use HeidelTime to extract dates from the associated articles. We than apply filtering and Gaussian fitting.

The parameters α, β and γ are optimised according to a Mean Distance Error (MDE) specifically tailored for temporal intervals (see Appendix A), which intuitively represents the percentage of overlap between the predicted intervals and the gold ones. For each approach we optimised the parameters α, β and γ by using an exhaustive GRID search on a randomly selected subset of 220 people.

3 Data

We applied the methodology on people's Wikipedia pages with the aim of measuring the performance of the proposed approaches. We define a person's temporal footprint as the time between their birth and death. This data has been selected in virtue of the availability of a vast amount of samples along with their curated lower and upper bounds, which are available through DBpedia (Auer et al., 2007). DBpedia was used to obtain a list of Wikipedia web pages about people born since 1000 AD along with their birth and death dates[1]. We checked the consistency of dates using some simple heuristics (the death date does not precede the birth date, a person age cannot be greater than 120 years) and discarded the incongruous entries. We collected 228,824 people who lived from 1000 to 2014. The Figure 2a shows the distribution of people according to the centuries, by considering people belonging to a particular century if they were born in it.

As input to our method, we used associated web pages with some sections discarded, typically containing temporal references invariably pointing to the present, such as *External links*, *See also*, *Citations*, *Footnotes*, *Notes*, *References*, *Further reading*, *Sources*, *Contents* and *Bibliography*. The majority of pages contains from 100 to 500 words (see Figure 2b).

4 Results

Figure 3 depicts the application of the proposed method to the Galileo Galilei's Wikipedia article. The aggregated results with respect to the MDE are showed in Table 1. The TEE Reg + Flt setting outperforms the other approaches. Still, the approaches that use the Gaussian fitting have lower standard deviation.

These results in Table 1 do not take into account the unbalance in the data due to the length of pages (the aggregate numbers are heavily unbalanced towards short pages i.e. those with less than 500 words, as depicted in Figure 2b). We therefore analysed the results with respect to the page length (see Figure 4). TEE RegEx method's performance is negatively affected by the length of the articles. The longer

[1]We used the data set `Persondata` and `Links-To-Wikipedia-Article` from DBpedia 3.9 (`http://wiki.dbpedia.org/Downloads39`)

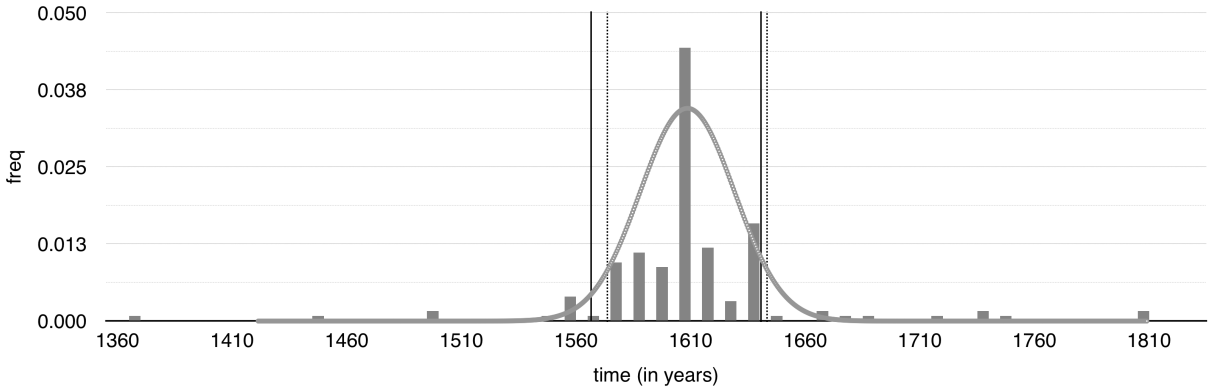


Figure 3: Graphical representation of the output on Galileo Galilei's Wikipedia page. Vertical continuous lines represent the prediction of temporal footprint boundaries, whereas dotted lines represent the real date of birth and death of the Italian scientist. The histogram shows the frequency of mentions of particular years in Galilei's Wikipedia page. The Gaussian bell is plotted in light grey.

Strategy	**Mean Distance Error**	**Standard Deviation**
TEE RegEx	0.2636	0.3409
TEE RegEx + Flt	**0.2596**	0.3090
TEE RegEx + Flt + FND	0.3503	0.2430
TEE Heidel + Flt + FND	0.5980	0.2470

Table 1: Results of the four proposed approaches.

a Wikipedia page is, the worse the prediction is. This is expected as longer articles are more likely to contain references to the past or future history, whereas in a short article the dates explicitly mentioned are often birth and death only. The use of the filter (*TEE RegEx+Flt*) generally improves the performance. The approaches that use the Gaussian fitting provide better results in case of longer texts. Still, in spite of its simplicity, the particular regular expression used in this experiment proved to be effective on Wikipedia pages and consequently an exceptionally difficult baseline to beat. Although counter-intuitive, *TEE RegEx + Flt + FND* performs slightly better than the HeidelTime-based method, suggesting that complex temporal information extraction systems do not bring much of useful mentions. This is in part due to the English Wikipedia's Manual of Style[2] which explicitly discourages authors from using implicit temporal expressions (e.g. *now*, *soon*, *currently*, *three years later*) or abbreviations (e.g. *'90*, *eighties* or *17th century*). Due to this bias, we expect a more positive contribution from using a temporal expression extraction system, when the methodology is applied on texts written without style constraints.

5 Conclusions

In this paper we introduced a method to extract temporal footprints of concepts based on mining their textual encyclopaedic description. The proposed methodology uses temporal expression extraction techniques, outlier filtering and Gaussian fitting. Our evaluation on people in Wikipedia showed encouraging results. We found that the use of a sophisticated temporal expression extraction system shows its strength only for long textual descriptions, whereas a simple regular expression-based approach performs better with short texts (the vast majority in Wikipedia pages).

The notion of temporal footprint has not to be interpreted strictly. A more factual interpretation of temporal footprint could be explored, such as temporal projection of a person's impact in history. This would allow to distinguish between people that made important contribution for the future history from those who did not. The predicted interval of Anna Frank's Wikipedia page is an

[2] `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style_(dates_and_numbers)#Chronological_items`

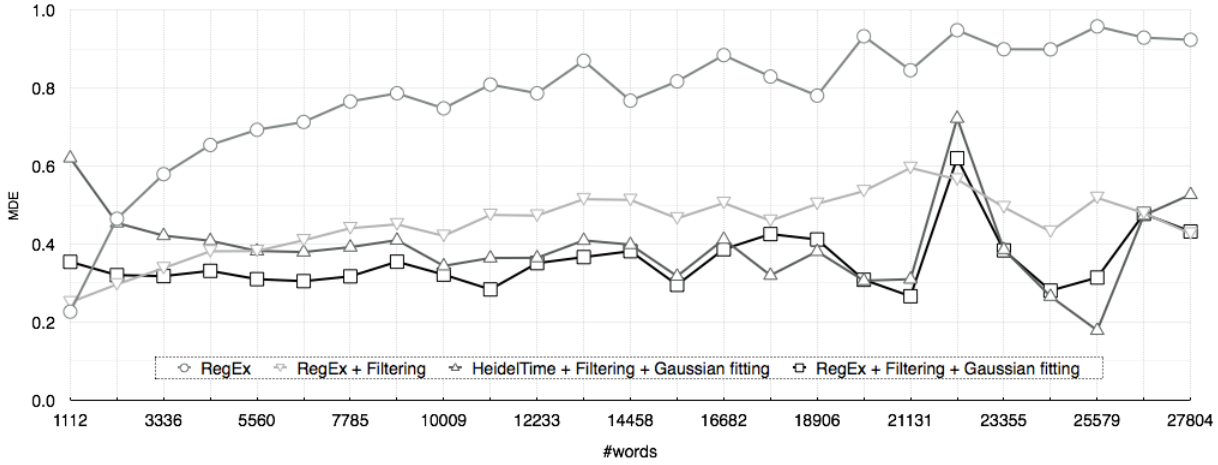


Figure 4: Observed error of the four proposed approaches with respect to the length of Wikipedia pages (the lower the better). Each data point represents the average of each bin. The *TEE RegEx* setting generally provide a very high error which is correlated with the page's length. The use of the outlier filter sensibly improves the performance (*TEE RegEx + Flt*). The approach *TEE RegEx + Flt + FND* is better than *TEE Heidel + Flt + FND* especially with short and medium size pages. The spike near 22000 words is due to a particular small sample.

example of that, and we invite the reader to investigate it via the online demo, which is available at: `http://www.cs.man.ac.uk/~filannim/projects/temporal_footprints/`. This site also provides the data, source code, optimisation details and supplementary graphs to aid the replicability of this work.

Acknowledgements

The authors would like to thank the reviewers for their comments. This paper has greatly benefited from their suggestions and insights. MF would like to acknowledge the support of the UK Engineering and Physical Science Research Council (EPSRC) in the form of doctoral training grant.

Appendix A: Error measure

In interval algebra, the difference between two intervals, $[A]$ and $[B]$, is defined as $[A] - [B] = [A_L - B_U, A_U - B_L]$ (where the subscripts $_L$ and $_U$ indicate lower and upper bound respectively). Unfortunately, this operation is not appropriate to define error measures, because it does not faithfully represent the concept of deviation (Palumbo and Lauro, 2003).

We therefore rely on distances for intervals, which objectively measure the dissimilarity between an observed interval and its forecast (Arroyo and Maté, 2006). In particular, we used De Carvalho's distance (De Carvalho, 1996):

$$d_{DC}([A],[B]) = \frac{d_{IY}^{\lambda}([A],[B])}{w([A] \cup [B])},$$

where $w([A] \cup [B])$ denotes the width of the union interval, and $d_{IY}^{\lambda}([A],[B])$ denotes the Ichino-Yaguchi's distance defined as follows:

$$d_{IY}^{\lambda}([A],[B]) = w([A] \cup [B]) - w([A] \cap [B]) + \lambda(2w([A] \cap [B]) - w([A]) - w([B])).$$

The Mean Distance Error (MDE) based on De Carvalho's distance is defined by:

$$MDE = \frac{1}{n}\sum_{t=1}^{n} \frac{d_{IY}^{\lambda=0}([A_t],[B_t])}{w([A_t] \cup [B_t])} = \frac{1}{n}\sum_{t=1}^{n} \frac{w([A_t] \cup [B_t]) - w([A_t] \cap [B_t])}{w([A_t] \cup [B_t])},$$

where n is the number of total samples. We set $\lambda = 0$ because we do not want to control the effects of the inner-side nearness and the outer-side nearness between the intervals.

The absence of any intersection between the intervals leads to the maximum error, regardless to the distance between the two intervals. A predicted interval far from the gold one has the same error of a predicted interval very close to the gold one, if they both not even minimally overlap with it.

References

Javier Arroyo and Carlos Maté. 2006. Introducing interval time series: Accuracy measures. *COMPSTAT 2006, proceedings in computational statistics*, pages 1139–1146.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Steven Bethard. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 10–14, Atlanta, Georgia, USA, June. Association for Computational Linguistics, Association for Computational Linguistics.

Fatima De Carvalho. 1996. Histogrammes et indices de proximité en analyse données symboliques. *Acyes de l'école d'été sur l'analyse des données symboliques. LISE-CEREMADE, Université de Paris IX Dauphine*, pages 101–127.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74, December.

Michele Filannino, Gavin Brown, and Goran Nenadic. 2013. ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 53–57, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Frank R Hampel. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1148–1158, Stroudsburg, PA, USA. Association for Computational Linguistics.

Immanuel Kant, Paul Guyer, and Allen W Wood. 1998. *Critique of pure reason*. Cambridge University Press.

Erdal Kuzey and Gerhard Weikum. 2012. Extraction of temporal facts and events from Wikipedia. In *Proceedings of the 2nd Temporal Web Analytics Workshop*, pages 25–32. ACM.

Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764 – 766.

Hector Llorens, Estela Saquete, and Borja Navarro. 2010. TIPSem (english and spanish): Evaluating CRFs and semantic roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291, Uppsala, Sweden, July. Association for Computational Linguistics.

Dat PT Nguyen, Yutaka Matsuo, and Mitsuru Ishizuka. 2007. Relation extraction from wikipedia using subtree mining. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, page 1414. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Francesco Palumbo and CarloN. Lauro. 2003. A PCA for interval-valued data based on midpoints and radii. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J.J. Meulman, editors, *New Developments in Psychometrics*, pages 641–648. Springer Japan.

Anisa Rula, Matteo Palmonari, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, Jens Lehmann, and Lorenz Bühmann. 2014. Hybrid acquisition of temporal scopes for rdf data. In *Proc. of the Extended Semantic Web Conference 2014*.

Jannik Strötgen, Julian Zell, and Michael Gertz. 2013. HeidelTime: Tuning english and developing spanish resources for tempeval-3. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 15–19, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Partha Pratim Talukdar, Derry Wijaya, and Tom Mitchell. 2012. Coupled temporal scoping of relational facts. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 73–82, New York, NY, USA. ACM.

Naushad UzZaman and James F. Allen. 2010. Event and temporal expression extraction from raw text: First step towards a temporally aware system. *International Journal of Semantic Computing*, 4(4):487–508.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 Task 15: TempEval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80, Prague.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 57–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fei Wu, Raphael Hoffmann, and Daniel S. Weld. 2008. Information extraction from Wikipedia: Moving down the long tail. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 731–739, New York, NY, USA. ACM.