iii the normalized TempEx's in their contexts,
iv the sentences containing the temponyms.

These features one-to-one correspond to the features we considered for computing the candidate mappings and their weights described in Section 5.1.

This task is realized by indexing all the features mentioned above using the Lucene search engine. For each temponym t of interest, we run a multi-field boolean search over the different features of the temponym, retrieving a set S_t of similar temponyms:

$$S_t = \{t' : sim_{Lucene}(t, t') \geq \tau\}$$

where sim_{Lucene} is the similarity score of the boolean vector space model provided by Lucene and τ is a specified threshold. Specifically, the similarity score is computed as:

$$sim_{Lucene}(t, t') = \sum_i \frac{v(t_i) \cdot v(t'_i)}{|v(t_i)||v(t'_i)|}$$

where t_i is the vector for feature group i (string, mentions, TempEx's, sentence) of temponym t.

For each temponym t the context features from the temponyms in S_t are merged. Thus, each temponym is enriched with the contextual information taken from highly similar temponyms in the corpus. Then, the ILP for the joint model is used to compute a solution for the global model. The data flow for the global model is illustrated in Figure 2.

6. EXPERIMENTS

In order to extensively evaluate our methods, we composed four hypotheses:

1. **Detection quality:** Our methods detect temponyms that are either events or facts with a significant coverage.
2. **Disambiguation quality:** Our methods significantly resolve the temponyms for different kinds of text.
3. **Temporal enrichment:** Our methods substantially add temporal information to documents by finding the temporal scopes of temponyms.
4. **Knowledge enrichment:** Our methods substantially add new knowledge to a knowledge base i) by finding new alias names for events and facts, and ii) by finding the time scope of knowledge base facts via anchoring them to resolved temponyms.

Since each hypothesis aims a different research goal, we developed different experimental settings to effectively assess each hypothesis independently. We first introduce the different datasets we used in the experiments.

6.1 Datasets

To evaluate the quality of our methods for temponym resolution, we performed experiments with three datasets with different characteristics: WikiWars, Biographies, and News.

WikiWars. The WikiWars corpus [28] has been popular in benchmarks for temporal tagging (i.e., resolving explicit, relative and implicit TempEx's). It contains 22 long Wikipedia articles about major wars in history. These articles are specifically rich in terms of TempEx's and named events. Thus, the temponyms detected in these articles are mostly of the event type. Note that WikiWars articles are plain text documents that do not contain any structured elements of Wikipedia such as entity links, categories, etc.

WikiBios. These are Wikipedia articles on the biographies of 30 prominent politicians (e.g., Barack Obama, Hugo Chávez, Vladimir Putin). We refer to this dataset as *WikiBios*. In contrast to the WikiWars, this corpus contains fewer event temponyms but features many temponyms that refer to temporal facts (awards, spouses, positions held, etc.). This makes it particularly challenging, since spotting facts is harder than spotting events which is a specific case of named entity disambiguation task. As WikiWars articles, WikiBios articles are plain text documents that do not contain any structured elements of Wikipedia. **News articles.** We show that our methods can perform well not only on properly edited texts that are rich in terms of events and facts (i.e., WikiWars, WikiBios) but also on the news that are compiled from a large source of news channels. We used GDELT (`http://gdeltproject.org/`) news dataset for our experiments. GDELT contains a set of entities for each article; however, we ignored these annotations and solely relied on our own methods to extract and disambiguate entities. In total, this test corpus contains 1,5 million news articles.

6.2 Evaluation Tasks and Metrics

To validate each hypothesis explained above we define an evaluation task.

Detection quality. We evaluated the quality of temponym detection by checking whether a detected noun phrase is indeed a temponym. We divided this task into two separate tasks: Event detection quality, and fact detection quality. For the event detection task, we manually annotated the named events appearing in WikiWars. In total, we annotated 1,154 events. We compare our method's coverage to the state-of-the-art entity disambiguation tool AIDA. In order to make a fair comparison in favor of AIDA, we only considered the named events that are linked to particular Wikipedia event articles by Wikipedia editors. Thus, we ended up 646 named event phrases with the respective sentences that they appear in.

For the fact detection task we manually annotated the facts appearing in WikiBios dataset. We only considered the first three paragraphs of each article during annotation. We annotated 589 temporal facts. The previous works [37, 40] consider only subject-verb-object style phrases for fact extraction. Since temponyms are of the noun phrase nature, we do not compare our method's coverage to previous work. Thus, we just report the recall values.

Disambiguation quality. The evaluation of mapping of temponyms is a human intelligence task. We evaluated the quality of temponym disambiguation by checking whether a temponym is mapped to the correct event or fact in the KB. This implies that the temporal scoping for the temponym is correct, too. We additionally checked whether the mentions in the temponym context are correctly disambiguated as well. There is no prior ground-truth for these corpora and creating such a dataset is a big amount of human work. Thus, we manually judged the quality of the computed mappings. We randomly selected 100 temponyms per model per dataset. In other words, 200 temponyms from WikiWars mappings, 300 from WikiBios mappings, and 300 from News mappings, a total of 800 temponym mappings. For statistical significance, we calculated Wilson confidence intervals [7].

We ran the local model, the joint model, and the global model on each corpus with the exception of WikiWars. The

	Strict			Relaxed		
Dataset	**Local**	**Joint**	**Global**	**Local**	**Joint**	**Global**
WikiBios	.54 ±.09	.60 ±.09	.68 ±.09	.61 ±.09	.66±.09	.76±.08
WikiWars	.75 ±.08	.82 ±.07	n/a	.84±.07	.86 ±.06	n/a
News	.58 ±.09	.64 ±.09	.67 ±.09	.69 ±.09	.75 ±.08	.79 ±.08

Table 5: Precision at 95% Wilson interval for different methods.

temponym	**Yago**	**Our model**	**Time scope**	**Eval**
the Great Recession	`GreatRecession`	`GreatRecession`	`[2007, 2009]`	Correct
the second term of Merkel	–	⟨`AngelaMerkel, holdsPosition, ChancellorOfGermany`⟩	`[2005, now]`	Okay
Obama's graduation	–	⟨`BarackObama, graduatedFrom, HarvardLawSchool`⟩	`[1991, 1991]`	Correct
the first Winter Olympics to be hosted by Russia	–	`2014WinterOlympics`	`[2014,2014]`	Correct
Putin's presidency	–	⟨`VladimirPutin, holdsPosition, PrimeMinisterOfRussia`⟩	`[2008, 2012]`	Wrong

Table 6: Example of temponyms mapped by our system vs Yago.

global model is not applicable here, as it requires multiple documents on the same or overlapping topics. In contrast, the 22 WikiWars articles are fairly disjoint in their contents and are not mentioned in GDELT news corpus much.

The evaluation is done by marking a mapping with three different scores; Correct, Okay, Wrong. Table 6 shows some examples of Correct, Okay, and Wrong matches. A mapping is considered "Okay" if it has partially correct match. For example, the temponym *the second term of Merkel* is mapped to the correct fact ⟨`AngelaMerkel, holdsPosition, ChancellorOfGermany`⟩ but it is marked as "Okay". The reason is that the second term of Angela Merkel is actually from 2009 to 2013 rather than from 2005 to now.

Precision is calculated in two different ways:

- For *strict* precision, we count the *Okay* mappings as wrong:
$$Precision_{strict} = \frac{\#Correct}{\#Total\ mappings}$$
- For *relaxed* precision, we count the *Okay* mappings as true:
$$Precision_{relaxed} = \frac{\#Correct\ +\ \#Okay}{\#Total\ mappings}$$

Temporal enrichment. To show our methods can substantially add extra temporal information to documents, we compare our methods to well known HeidelTime tagger by running the both methods on WikiWars and WikiBios datasets. We compare the number of normalized TempEx's by HeidelTime tagger to the number of normalized temponyms by our methods.

Knowledge enrichment. The temponym resolution task has two important outcomes in terms of knowledge enrichment: First, temponym resolution enriches the KB by providing additional paraphrases for known events and facts. For example, our methods can add the temponym *"the largest naval battle in history"* as an alias for the event `BattleOfLeyteGulf`, or *"Obama's presidency"* as an alias for the fact `BarackObama, holdsPosition, presidentOfUS`. We add this new knowledge to the KB through "rdfs:label" triples that have a temponym phrase as subject, and an event or a fact identifier as object. We call this task *Knowledge paraphrasing*.

We assess the knowledge paraphrasing, by comparing outcome of our methods to Yago2 knowledge base in terms of paraphrase coverage. Therefore, we randomly chose 100 correctly mapped temponyms and checked how many temponyms are already known to Yago2, either as an event entity or as a fact. We built a text index over all events and facts in Yago2 and their alias names. For the randomly chosen 100 temponyms, we queried this index for each temponym and took the top-10 most relevant results for each query. We manually checked all these returned answers, thus considering also approximate matches for a fair comparison in favor of Yago2.

Second, temponym resolution also enhances the fact extraction tools for knowledge bases by providing them additional temporal and semantic clues. For example, in the sentence "Ronaldo joined Real Madrid during second term of Florentino Pèrez" a fact extraction tool can extract the fact ⟨`f1:CristianoRonaldo, playsFor, RealMadrid`⟩ but no time scope attached. Temponym resolution would normalize the phrase *second term of Florentino Pèrez* to time [2009, now] by mapping it to the fact ⟨`f2:FlorentinoPèrez, isPresidentOf, RealMadrid, [2009, now]`⟩. Thus, a fact extraction tool can temporally link two facts as a new fact ⟨`f3:f1, validDuring, f2`⟩. We call this task *Knowledge linking*.

For the knowledge linking task, we carried out an extrinsic case study. We modified the PATTY's binary fact extraction patterns to ternary patterns so that they can take a temponym as an argument. For example, the PATTY pattern ⟨*subject, verb, object*⟩ is modified to ⟨*subject, verb, object, preposition, temponym*⟩. Thus, a fact extracted from ⟨*subject, verb, object*⟩ triple can be linked to the particular temponym through a particular preposition such as "during, before, after". For this task, we ran PATTY tool on its extraction corpus. We report the number of facts that

	WikiWars	WikiBios
# Gold annotations	646	589
# AIDA's extractions	186	–
# Our extractions	338	194
AIDA's recall	29%	–
Our recall	52%	33%

Table 7: Recall values for AIDA and for our method.

are linked to temponyms through three prepositions "during, before, after".

6.3 Results

Detection quality. Our methods detected 233 265 temponyms from the three corpora. Specifically, 2 504 temponyms from WikiWars, 5 390 from WikiBios, and 225 371 from the News dataset are extracted. The recall values we calculated specifically for events in WikiWars and for facts in WikiBios datasets are shown in Table 7.

i) Event detection. Among the 646 annotated named events in WikiWars dataset, AIDA detected 186 of them, which results in 29% coverage. On the contrary, our methods detected 338 of the events, which resulted in 52% coverage. It is obvious that general ail NED tools such as AIDA are not well suited for event detection. Therefore, specialized solutions such as our methods should be pursued.

ii) Fact detection. Among the 589 annotated temporal facts in WikiBios dataset, our method detected 194 of them, which yields a 33% coverage. It might seem a low coverage. However, considering that temporal facts can be phrased in text in many different ways, our results are encouraging. Our empirical observations show that the main cause of the low coverage is the deficacy of the KB. Using a larger knowledge base may improve the results. Secondly, enlarging the pattern dictionary might have a direct impact on the coverage.

Disambiguation quality. We evaluated the overall disambiguation quality over randomly selected 800 temponym mappings. We computed 95% Wilson confidence intervals for strict precision and for relaxed precision, on all three datasets. The strict matching evaluation gives us a 65% $\pm$ 0.03 precision. The relaxed matching evaluation gives us a 73% $\pm$ 0.03. The detailed precision results for each dataset and for each method are shown in Table 5.

We see that the joint and global models boost the precision by a large margin. For the relaxed precision measure, the global models achieved substantial gains over the joint models. The precision numbers are particularly good for the News and the WikiWars corpora, thus achieving high value for semantic markup and knowledge enrichment. For WikiBios, the results are somewhat worse. Here we faced the challenge that many temponyms refer to SPOT facts (e.g., awards, spouses, children, held positions, etc.) rather than typed events, which is much harder to deal with. Nevertheless, the results are very encouraging, given that temponym resolution is more demanding than TempEx resolution and the state-of-the-art results for TempEx's are 80 to 90% [38].

Temporal enrichment. We compared our best performing model, global model, to HeidelTime tagger to see how much additional temporal information is added to documents. HeidelTime normalized 5 533 TempEx's from WikiBios dataset, and 2 047 from WikiWars dataset to date values. Whereas, our methods normalized 885 temponyms from WikiBios dataset, and 558 from WikiWars dataset to date values by disambiguating these temponyms to KB facts or events. Note that these temponyms are not detected by HeidelTime tagger at all. Thus, our methods add 16% additional temporal information to WikiBios dataset and 27% to WikiWars dataset.

Knowledge enrichment. For the knowledge paraphrasing task, the manual assessment over randomly selected 100 temponyms showed that Yago2 knows alias names for only 52 of the events given by the 100 temponyms. On the remaining 48, Yago2 does not even have any approximate matches. Yago2's coverage is great for canonicalized event names such as "the Great Recession", "Second World War", etc. However, it is largely agnostic to phrases for less standardized events such as "the second term of Merkel", "Obama's graduation", "the last presidential election in France", etc. Our methods do not only detect these temponyms but also disambiguate them correctly onto events or facts. Examples from this comparison are shown in Table 6.

For the knowledge linking task, our methods disambiguated 65 625 temponyms surrounding the facts that are extracted by ternary patterns. 12 803 (20%) of these temponyms are temporally linked to the extracted facts through prepositional links. For example, the base facts extracted from the sentence "Hillary was First Lady of the United States during Clinton's tenure." by this method are

⟨`f1:HillaryClinton, holdsPosition, FirstLadyOfUS`⟩,
⟨`f2:BillClinton, holdsPosition, PresidentOfUS`⟩.

These two base facts are linked through the reification mechanism of RDFS. Thus, `f1` and `f2` are linked as

⟨`f3:f1, validDuring, f2`⟩.

6.4 Data and Software

The data used for our experiments is publicly available.[2] Moreover, we incorporated some of our methods into the well known temporal tagger HeidelTime. Further information how to use this new version of HeidelTime can be obtained from the same URL.

7. CONCLUSION

We have presented a viable solution for temponym resolution – an important problem for search, text analytics and KB curation that has received little attention in the literature so far. Our experiments demonstrate that we can resolve temponyms onto events or facts in a KB with fairly good precision, and that we can enrich the KB itself with additional names for known events and with newly emerging events. Our future work includes scaling our system up for processing very large text corpora, testing our methods with different knowledge bases and with a larger pattern dictionary. We expect that the semantic markup of temponyms in news articles and social media will boost next-generation deep analytics of unstructured data.

8. REFERENCES

[1] O. Alonso, M. Gertz, and R. Baeza-Yates. Enhancing Document Snippets Using Temporal Information. In *SPIRE*, 2011.

[2] `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/evin/`

[2] O. Alonso and K. Khandelwal. Kondenzer: Exploration and Visualization of Archived Social Media. In *ICDE*, 2014.

[3] O. Alonso, J. Strötgen, R. Baeza-Yates, and M. Gertz. Temporal Information Retrieval: Challenges and Opportunities. In *TWAW*, 2011.

[4] A. Angel, N. Sarkas, N. Koudas, and D. Srivastava. Dense Subgraph Maintenance Under Streaming Edge Weight Updates for Real-time Story Identification. *Proc. VLDB Endow.*, 5(6):574–585, 2012.

[5] K. Berberich, S. J. Bedathur, O. Alonso, and G. Weikum. A Language Modeling Approach for Temporal Information Needs. In *ECIR*, 2010.

[6] S. Bethard and J. H. Martin. Identification of Event Mentions and Their Semantic Class. In *EMNLP*, 2006.

[7] L. D. Brown, T. T. Cai, and A. Dasgupta. Interval Estimation for a Binomial Proportion. *Statistical Science*, 16:101–133, 2001.

[8] D. Carmel, M.-W. Chang, E. Gabrilovich, B.-J. P. Hsu, and K. Wang. First International Workshop on Entity Recognition & Disambiguation. 2014.

[9] A. Chang and C. Manning. SUTime: A Library for Recognizing and Normalizing Time Expressions. In *LREC*, 2012.

[10] P.-T. Chang, Y.-C. Huang, C.-L. Yang, S.-D. Lin, and P.-J. Cheng. Learning-based Time-sensitive Re-ranking for Web Search. In *SIGIR*, 2012.

[11] M. Cornolti, P. Ferragina, and M. Ciaramita. A Framework for Benchmarking Entity-annotation Systems. In *WWW*, 2013.

[12] A. Das Sarma, A. Jain, and C. Yu. Dynamic Relationship and Event Discovery. In *WSDM*, 2011.

[13] Q. Do, W. Lu, and D. Roth. Joint Inference for Event Timeline Construction. In *EMNLP-CoNLL*, 2012.

[14] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[15] J. R. Finkel, T. Grenager, and C. Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *ACL*, 2005.

[16] D. Gupta and K. Berberich. Identifying Time Intervals of Interest to Queries. In *CIKM*, 2014.

[17] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artif. Intell.*, 194:28–61, 2013.

[18] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust Disambiguation of Named Entities in Text. In *EMNLP*, 2011.

[19] P. Jindal and D. Roth. Extraction of Events and Temporal Expressions from Clinical Narratives. *Journal of Biomedical Information*, 46:S13–S19, 2013.

[20] N. Kanhabua, T. Ngoc Nguyen, and W. Nejdl. Learning to Detect Event-Related Queries for Web Search. In *WWW*, 2015.

[21] Z. Kozareva and E. H. Hovy. Learning Temporal Information for States and Events. In *ICSC*, 2011.

[22] E. Kuzey and G. Weikum. Extraction of Temporal Facts and Events from Wikipedia. In *TempWeb*, 2012.

[23] K. Lee, Y. Artzi, J. Dodge, and L. Zettlemoyer. Context-dependent Semantic Parsing for Time Expressions. In *ACL*, 2014.

[24] J. L. Leidner. Toponym Resolution in Text. *SIGIR Forum*, 41(2):124–126, 2007.

[25] X. Li and W. B. Croft. Time-based Language Models. In *CIKM*, 2003.

[26] M. D. Lieberman and H. Samet. Adaptive Context Features for Toponym Resolution in Streaming News. In *SIGIR*, 2012.

[27] X. Ling and D. S. Weld. Temporal Information Extraction. In *AAAI*, 2010.

[28] P. Mazur and R. Dale. WikiWars: A New Corpus for Research on Temporal Expressions. In *EMNLP*, 2010.

[29] D. Metzler, R. Jones, F. Peng, and R. Zhang. Improving Search Relevance for Implicitly Temporal Queries. In *SIGIR*, 2009.

[30] N. Nakashole, G. Weikum, and F. Suchanek. PATTY: A Taxonomy of Relational Patterns with Semantic Types. In *EMNLP-CoNLL*, 2012.

[31] J. Piskorski, J. Belayeva, and M. Atkinson. Exploring the Usefulness of Cross-lingual Information Fusion for Refining Real-time News Event Extraction: A Preliminary Study. In *RANLP*, 2011.

[32] J. Pustejovsky, K. Lee, H. Bunt, and L. Romary. ISO-TimeML: An International Standard for Semantic Annotation. In *LREC*, 2010.

[33] D. Shahaf and C. Guestrin. Connecting the Dots Between News Articles. In *KDD*, 2010.

[34] W. Shen, J. Wang, and J. Han. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *Trans. on Knowl. and Data Eng.*, 27(2):443–460, 2015.

[35] M. Spaniol, J. Masanès, and R. Baeza-Yates. The 4th Temporal Web Analytics Workshop. 2014.

[36] J. Strötgen and M. Gertz. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2):269–298, 2013.

[37] P. P. Talukdar, D. Wijaya, and T. Mitchell. Coupled Temporal Scoping of Relational Facts. In *WSDM*, 2012.

[38] N. UzZaman, H. Llorens, L. Derczynski, J. F. Allen, M. Verhagen, and J. Pustejovsky. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *SemEval*, 2013.

[39] M. Verhagen and J. Pustejovsky. The TARSQI Toolkit. In *LREC*, 2012.

[40] Y. Wang, M. Dylla, M. Spaniol, and G. Weikum. Coupling Label Propagation and Constraints for Temporal Fact Extraction. In *ACL*, 2012.

[41] Y. Wang, B. Yang, L. Qu, M. Spaniol, and G. Weikum. Harvesting Facts from Textual Web Sources by Constrained Label Propagation. In *CIKM*, 2011.

[42] S. Whiting, J. Jose, and O. Alonso. Wikipedia As a Time Machine. In *WWW*, 2014.

[43] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang. Evolutionary Timeline Summarization: A Balanced Optimization Framework via Iterative Substitution. In *SIGIR*, 2011.

[44] X. Zhao, Y. Guo, R. Yan, Y. He, and X. Li. Timeline Generation with Social Attention. In *SIGIR*, 2013.