Temponym Tagging: Temporal Scopes for Textual Phrases

Erdal Kuzey, Jannik Strötgen, Vinay Setty, Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
{ekuzey,jannik.stroetgen,vsetty,weikum}@mpi-inf.mpg.de

ABSTRACT

For many NLP and IR applications, anchored temporal information extracted from textual documents is of utmost importance. Thus, temporal tagging – the extraction and normalization of temporal expressions – has gained a lot of attention in recent years and several tools such as HeidelTime and SUTime are proposed. However, such tools do not address textual phrases with temporal scopes like "Clinton's time as First Lady". While such phrases (so-called *temponyms*) are not temporal expressions per se, information about their temporal scopes can be helpful in many scenarios, e.g., in the context of temporal information retrieval. In this paper, we describe the integration of a wide range of temponyms to the publicly available temporal tagger HeidelTime to include temponym tagging.

Keywords

temporal tagging; temporal scopes; temponyms; HeidelTime

1. INTRODUCTION & DEFINITION

The value of temporal information for many tasks has been widely discussed, e.g., in the context of information retrieval [1, 2]. To extract temporal information latently available in the documents' content, temporal taggers can be applied. They detect temporal expressions and make their semantics accessible by normalizing them into a standard format. In addition to explicit (*\April 2016"*, 2016-04) and implicit expressions (*\Christmas 2015"*, 2015-12-25), underspecified (*\in April"*) and relative expressions (*\two days later"*) are also covered. To normalize these expressions, temporal taggers typically detect reference times and resolve temporal relations to them [7]. For instance, assuming April 11, 2016 is detected as reference date for the expression *\two days later"*, the latter can be normalized to 2016-04-13.

State-of-the-art temporal taggers like SUTime [3] and HeidelTime [8] follow the TimeML standard[1] and assign TIMEX3 tags to date, time, duration, and set expressions. However, in addition to such expressions, there are many textual phrases with temporal scopes although they are not temporal expressions per se. For example, given the text *\The work is presented at the WWW Conference 2016"*, a temporal tagger detects *\2016"* (2016), although *\WWW Conference 2016"* refers to [2016-04-11,2016-04-15].

Recently, the concept of temponyms was defined in [5]:

> "Free-text temporal expressions refer to arbitrary kinds of named events or facts with temporal scopes that are merely given by a text phrase but have unique interpretations given the context and background knowledge."

For the detection and resolution of temponyms, an entity-time oriented document model relying on the Yago knowledge base for distant supervision was presented in addition to the concept of temponyms [5]. Multiple contextual cues such as temporal expressions, part-of-speech tags, and entity mentions are used to jointly disambiguate temponym mentions and to infer their temporal scopes by matching the events and facts to entries in Yago. Thus, this approach can handle explicit (*\Clinton's presidency"*) and ambiguous temponyms such as *\his presidency"* in texts like *\During his presidency, Clinton . . . "*. However, due to relying on a wide set of tools and features as well as a suit of Integer Linear Programs, scaling up the system for processing large text corpora was left as an open issue.

In this paper, we extend the publicly available temporal tagger HeidelTime[2] to cover a subset of temponyms, namely explicit temponyms that are unambiguous even without a context. While we address event-style (*\FIFA World Cup Final 1998", 1998-07-12)* and fact-style temponyms (*\Clinton's term as secretary of state"*, [2009-01-21,2013-02-01]), we do not address temponyms for which rich preprocessing is required, e.g., coreference resolution for temponyms such as *\his presidency"*. Thus, we do not need a deep analysis and disambiguation of textual phrases. In addition, our approach can run stand-alone without distant supervision.

While temporal taggers already enrich documents with valuable semantics, the extraction of temponyms brings temporal information to the next level by making further information available and by specifying it in a more fine-grained manner. Our approach is thus valuable for many search and exploration tasks.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889289.

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

[2] https://github.com/HeidelTime/heideltime

2. TEMPONYM TAGGING

2.1 Temponym Creation Process

We combine three data sources to create a large collection of explicit temponyms together with their temporal scopes.
i) **Yago**[3] provides temporal predicates such as `holdsPosition` and `isMarriedTo`. The facts of such predicates have a time scope attached. Moreover, Yago contains the predicates `startedOnDate`, `endedOnDate`, `happenedOnDate` indicating the start and end dates of events.
ii) **Aida** [4] contains alias names for entities. For example, *\London Olympic Games", \Games of the XXX Olympiad"* are alias names for the entity *h*`2012 Summer Olympics`*i*.
iii) The **temponym pattern dictionary** created in [5] contains noun phrases mapped to Yago's temporal predicates.

For facts, we create temponyms in two steps: First, all the temporal facts from Yago are paraphrased using alias names from the Aida dictionary, e.g., a paraphrase of *h*`Hillary Rodham Clinton holdsPosition United States Secretary of State`*i* is *h*`Hillary Clinton holdsPosition Secretary of State`*i*. Then, temponyms are created by mapping the predicates of these facts to the noun phrases in the pattern dictionary. Therefore, the paraphrased fact *h*`Hillary Clinton holdsPosition Secretary of State`*i* is converted to the temponym *\Hillary Clinton's term as US Secretary of State"* by mapping the predicate `holdsPosition` to patterns such as *\[PER]'s term as", \[PER] serving as"*, etc.

For events, we create temponyms by taking unique alias names for the events from the Aida dictionary. For example, *h*`2012 Summer Olympics`*i* event has *\London Olympic Games", \Games of the XXX Olympiad"* as temponyms together with the canonical name of the event itself.

2.2 Temponym Tagging with HeidelTime

HeidelTime is a rule-based, domain-sensitive, and multilingual temporal tagger, which contains language-dependent resources. *Pattern resources* contain phrases that may be matched (e.g., "April"), *normalization resources* contain information how to normalize them (e.g., *\April"* to `04`), and *rules* handle the extraction and normalization jointly.

In the extraction parts of the rules, the pattern resources can be accessed and regular expressions and part-of-speech constraints can be defined. In the normalization parts, it is defined how matched patterns can be resolved by accessing normalization resources. A simple rule is *ext=\%month*$_1$ *of %year*$_2$*", norm=\normYear(2)-normMonth(1)"*, e.g., to normalize phrases such as *\April of 2002"* to `2002-04`. In addition, context-dependent and domain-dependent constraints might be defined, but as our approach to temponym resolution does not cover ambiguous expressions, we refer for further details about HeidelTime's rule syntax to [7, 8].

For temponym tagging, we use the temponyms described in Section 2.1 to create HeidelTime pattern and normalization resources for several temponym subtypes (e.g., event, fact-birth). Simple rules combine this information and are applicable in the same way as all rules for standard temporal expressions. Verbal paraphrases are also added. For instance, using the patterns `deadPerson` (names of persons for who dates of death are available) and `deadVerb` (verbal phrases such as *\died"* and *\was killed"*), the rule *ext='%deadPerson*$_1$ *%deadVerb', norm='normDeadPerson(1)'* matches phrases about a person's death and assigns respective dates.

3. STATISTICS & EXPERIMENTS

Overall, we added to HeidelTime's English resources temponyms for more than 40,000 events and for more than 900,000 facts with several paraphrases. Currently, facts about persons (birth, death, political position, marriage; 664,000) and culture (releases, directions, authors; 253,000) are contained. The events cover several types, e.g., sport events and (historic) battles. Note that we only added those temponyms, for which explicit start and end information is available in Yago. We are currently working on improving the temporal scopes of a second set of about 35,000 events.

To test how many temponyms are detected and normalized on an example corpus, we ran HeidelTime on the WikiWars corpus [6], which contains 22 Wikipedia articles about important wars in history with 2,681 manually annotated standard temporal expressions. We extracted a total of 212 temponyms so that about 8% additional temporally annotated phrases are detected in addition to the standard temporal expressions. Since all extracted temponyms are explicit while many of the standard temporal expressions are not explicit and thus difficult to normalize (cf. [7]), temponyms become even more valuable. One issue, however, is that many further phrases in WikiWars could also be associated with temporal scopes if not only explicit temponyms were tackled, e.g., by approaches as the one introduced in [5].

4. CONCLUSIONS

We presented a HeidelTime extension to extract and normalize *temponyms*, free-text phrases with temporal scopes. So far, we only addressed explicit temponyms. However, we believe that by integrating further subtypes of temponyms, adding more paraphrases, and including temporal scopes currently not part of Yago, the value of temponym tagging can be further improved. In addition, we plan to use Yago's Wikipedia identifiers and Wikipedia's inter-language links to add temponyms for further languages to HeidelTime.

5. REFERENCES

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

[2] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt. Survey of Temporal Information Retrieval and Related Applications. *ACM Computing Surveys*, 47(2), 2014.

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

[4] 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.

[5] E. Kuzey, V. Setty, J. Strötgen, and G. Weikum. As Time Goes By: Comprehensive Tagging of Textual Phrases with Temporal Scopes. In *WWW*, 2016.

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

[7] J. Strötgen. *Domain-sensitive Temporal Tagging for Event-centric Information Retrieval*. PhD thesis, Heidelberg University, 2015.

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

[3] http://www.yago-knowledge.org