HeidelTime: Tuning English and Developing Spanish Resources for TempEval-3

Jannik Strötgen Julian Zell Michael Gertz
Institute of Computer Science, Heidelberg University
Im Neuenheimer Feld 348, 69120 Heidelberg, Germany
{stroetgen,gertz}@uni-hd.de, j.zell@stud.uni-heidelberg.de

Abstract

In this paper, we describe our participation in the TempEval-3 challenge. With our multilingual temporal tagger HeidelTime, we addressed task A, the extraction and normalization of temporal expressions for English and Spanish. Exploiting HeidelTime's strict separation between source code and language-dependent parts, we tuned HeidelTime's existing English resources and developed new Spanish resources. For both languages, we achieved the best results among all participants for task A, the combination of extraction and normalization. Both the improved English and the new Spanish resources are publicly available with HeidelTime.

1 Introduction

The task of temporal annotation, which is addressed in the TempEval-3 challenge, consists of three subtasks: (A) the extraction and normalization of temporal expressions, (B) event extraction, and (C) the annotation of temporal relations (UzZaman et al., 2012). This makes sub-task A, i.e., temporal tagging, a prerequisite for the full task of temporal annotating documents. In addition, temporal tagging is important for many further natural language processing and understanding tasks, and can also be exploited for search and exploration scenarios in information retrieval (Alonso et al., 2011).

In the context of the TempEval-2 challenge (Verhagen et al., 2010), we developed our temporal tagger HeidelTime (Strötgen and Gertz, 2010), which achieved the best results for the extraction and normalization of temporal expressions for English documents. For our work on multilingual information retrieval (e.g., Strötgen et al. (2011)), we extended HeidelTime with a focus on supporting the simple integration of further languages (Strötgen and Gertz, 2012a). For TempEval-3, we now tuned HeidelTime's English resources and developed new Spanish resources to address both languages that are part of TempEval-3. As the evaluation results demonstrate, HeidelTime outperforms the systems of all other participants for the full task of temporal tagging by achieving high quality results for the extraction and normalization for English and Spanish.

The remainder of the paper is structured as follows: We explain HeidelTime's system architecture in Section 2. Section 3 covers the tuning of HeidelTime's English and the development of the Spanish resources. Finally, we discuss the evaluation results in Section 4, and conclude the paper in Section 5.

2 HeidelTime

HeidelTime is a multilingual, cross-domain temporal tagger. So far, it can process English, German, and Dutch text. In previous work, we analyzed domain-dependent challenges and demonstrated that domain-sensitive strategies for normalizing temporal expressions result in significant normalization improvements when switching between news- and narrative-style documents (Strötgen and Gertz, 2012b). Although TempEval-3 only addresses news documents, the tuned English and new Spanish resources can be used to process news and also narrative-style documents such as Wikipedia articles with high extraction and normalization quality.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 15–19, Atlanta, Georgia, June 14-15, 2013.

Architecture of HeidelTime. HeidelTime is a rule-based system with a strict separation between source code and language-dependent resources. While the strategies for processing different domains are part of the source code, resources consist of files for (i) patterns, (ii) normalizations, and (iii) rules. They are read by HeidelTime's resource interpreter and thus have to be developed based on HeidelTime's well-defined rule syntax.

The pattern files contain words and phrases, which are typically used to express temporal expressions, e.g., names of months. The normalization files contain normalization information about the patterns, e.g., the value of a specific month's name. Finally, the rule files contain rules for date, time, duration, and set expressions.

All rules have an extraction part and a normalization part. The extraction part, in which the pattern resources can be used for generalization, defines the expressions that have to be matched in a document. The normalization part normalizes the context-independent content of the expression using the normalization resources. While explicit temporal expressions (e.g., *May 1st, 2013*) can directly be fully normalized, underspecified (*November*) and relative (*today*, *two weeks ago*) expressions can only be normalized in an underspecified manner. The full normalization depends on the domain of the document that is to be processed and the context of the expression. For this, HeidelTime applies domain-sensitive strategies to normalize such expressions during its disambiguation phase, which is called after the extraction and the normalization phases.

The TempEval-3 data is from the news domain. Here, HeidelTime usually uses the document creation time as reference time. The temporal relation to it is identified based on the tense in the sentence.[1]

Preprocessing. HeidelTime requires sentence, token, and part-of-speech information. For this, the TreeTagger (Schmid, 1994) is used. Since there is a Spanish model for the TreeTagger, adding Spanish preprocessing capabilities to HeidelTime was fairly easy. A wrapper for the TreeTagger is also part of the UIMA HeidelTime kit described next.

[1]For further details on HeidelTime's rule syntax, its domain-dependent normalization strategies, and its architecture in general, we refer to Strötgen and Gertz (2012a).

UIMA HeidelTime kit. For processing TempEval-3 data, we used the UIMA version of HeidelTime, developed a collection reader and a CAS consumer to read and write TempEval-3 input and output data, and added both components to our UIMA HeidelTime kit. This makes HeidelTime's evaluation results reproducible on the training and test sets.

3 HeidelTime for TempEval-3

In TempEval-3, we participated with one Spanish and three English runs: For Spanish, we used our newly developed resources. For English, we used (i) HeidelTime 1.2, which was released in May 2012, (ii) a version containing several bug fixes and improvements, which were implemented independently from TempEval-3, and (iii) HeidelTime with its new English resources tuned for TempEval-3.

In general, our goal when developing HeidelTime resources is to achieve high quality normalization results. Thus, we only want to extract temporal expressions which can be normalized correctly with high probability – an issue, which will be further looked at in the discussion in the evaluation section. Before that, we next describe language-independent adaptations to HeidelTime. Then, we present the tuning of the English resources (Section 3.2) and the development of the Spanish resources (Section 3.3).

3.1 General HeidelTime Adaptations

We performed the following language-independent changes to HeidelTime:

(i) Weekday normalization: In news-style documents, extracted weekdays that are equal to the weekday of the document creation time (dct) are now normalized to the date of the dct independent of the tense in the sentence.

(ii) Century/decade normalization: So far, decade and century expressions were not correctly normalized by HeidelTime according to TimeML, e.g., "199X" instead of "199" for "the 1990s".

The first change is based on the intuitive assumption that information in news-style documents is temporally focused around the dct. In addition, this assumption is supported by the English and the Spanish training data. The second change is related to the annotation standard. Both changes can thus be generalized in a language-independent manner.

3.2 Tuning HeidelTime's English Resources

Three training corpora were provided by the organizers: the Aquaint and TimeBank gold standard corpora, and a large corpus referred to as silver standard, which was created by merging results of three tools (Llorens et al., 2012). After a brief analysis, we decided not to use the silver standard due to the rather low annotation quality. Motivated by observations in the gold standard corpora, we performed the following English-specific modifications in addition to the general adaptations described above:

(i) REF-value expressions: expressions normalized to past, present, or future are not consistently annotated in the training data. Since such expressions are rather less valuable for further tasks and to avoid false positives, we removed some of those patterns from the resources.

(ii) Ambiguous expressions: We added negative rules for expressions such as *may*, *march*, and *fall* to filter them out if they do not refer to a date.

(iii) Article/modifier: We allowed some more combinations of articles and modifiers.

Note that HeidelTime was already a state-of-the-art tool for English temporal tagging so that the changes are rather minor.

3.3 Developing Spanish Resources

In this section, we explain the resource development process for Spanish. Then, we detail language-specific challenges we faced during this process.

Resource Development Process. So far, there were no HeidelTime resources for Spanish, and we thus started the development from scratch.

(i) Preprocessing: As mentioned in Section 2, we use the TreeTagger with its Spanish module for sentence, token, and part-of-speech annotation.

(ii) Translation of pattern files: Starting with HeidelTime's English pattern resources, we developed the Spanish pattern resources. The goal was that all patterns that are frequently used to express temporal expressions are included in the resources. Note that it is not important that the patterns are context independent. The context in which a pattern should occur can be defined within the rules.

(iii) Translation of normalization files: Similar to the patterns, we translated the English normalization files and adapted them to the new Spanish patterns.

(iv) Rule Development: Based on the English rules for dates, times, durations, and sets, we developed similar Spanish rules. Using the Spanish training corpus to check for partially matching patterns, false positives, false negatives, and incorrect normalizations, we then iteratively adapted the rules, but also the pattern and normalization resources.

Challenges. Spanish as a Romance language is rich in inflection. Nouns, adjectives, and determiners are inflected with respect to number and gender. During the development of the pattern and normalization resources, this had to be taken into account.

As for nouns, there are many inflection forms of verbs in Spanish, e.g., to represent tense. While verbs are usually not part of temporal expressions, the inflection of verbs has to be considered for the normalization of ambiguous expressions such as *el lunes* (Monday) or *junio* (June). As mentioned above, in news-style documents, HeidelTime uses the tense of the sentence to determine the relation to the reference time, i.e., to decide whether the expression refers to a previous or upcoming date.

The tense is determined using part-of-speech information, and, if necessary, pattern information of words with specific part-of-speech tags. For each language, this information is defined in the pattern resources. Unfortunately, the Spanish tag-set of the TreeTagger module does not contain tags covering tense information, e.g., all finite lexical verbs are tagged as VLfin. Thus, we created regular expression patterns to match typical inflection patterns representing tense information and check words tagged as verbs by the tagger for these patterns.

However, due to the ambiguity of the Spanish inflection, we can only add patterns to detect future tense. If no tense is identified, the year is set to the year of the reference time. As detailed in the discussion of the evaluation results described in Section 4, identifying the correct relation to the reference time is a frequent source of normalization errors.

4 Evaluation Results

Measures. For the extraction task, precision (P), recall (R), and f_1-score (F1) are used for strict and relaxed matching. The value F1 and type F1 measures combine relaxed matching with correct normalization. Systems are ranked by value F1 (value).

	strict match			relaxed match			normalization	
a) Aquaint	P	R	F1	P	R	F1	**value**	type
tuned	80.17	81.69	80.92	90.85	92.57	91.7	72.37	83.32
bug-fixed	77.56	81.17	79.32	88.28	92.40	90.30	70.21	82.03
1.2	73.32	81.17	77.05	83.46	92.40	87.70	67.87	79.67
b) TimeBank	P	R	F1	P	R	F1	**value**	type
tuned	85.39	84.15	84.76	92.16	90.83	91.49	79.01	88.74
bug-fixed	83.17	82.70	82.94	90.86	90.35	90.60	76.24	87.78
1.2	82.89	82.62	82.76	90.72	90.43	90.57	76.39	87.75
c) Spanish	P	R	F1	P	R	F1	**value**	type
new	90.53	81.26	85.65	96.23	86.38	91.04	84.10	89.40

Table 1: Results on training data ranked by *value F1*.

	strict match			relaxed match			normalization	
a) English	P	R	F1	P	R	F1	**value**	type
tuned	83.85	78.99	81.34	93.08	87.68	90.30	**77.61**	82.09
bug-fixed	80.77	76.09	78.36	90.00	84.78	87.31	72.39	79.10
1.2	80.15	76.09	78.07	89.31	84.78	86.99	72.12	78.81
next best*	78.72	80.43	79.57	89.36	91.30	90.32	70.97	80.29
b) Spanish	P	R	F1	P	R	F1	**value**	type
HeidelTime	90.91	80.40	85.33	96.02	84.92	90.13	**85.33**	87.47
TipSemB	88.51	77.39	82.57	93.68	81.91	87.40	71.85	82.04
jrc-1/2	65.83	39.70	49.53	86.67	52.26	65.20	50.78	62.70

Table 2: TempEval-3 task A evaluation results ranked by *value F1* (* next best: NavyTime).

Results on Training Data. Table 1 shows the results on the Aquaint (a), TimeBank (b), and Spanish training corpora (c). On both English corpora, HeidelTime's TempEval-3 tuned version outperforms the other two versions. The big differences between the two English corpora are rather due to the better annotation quality of TimeBank than due to different challenges in the documents of the two corpora.

TempEval-3 Evaluation. The evaluation results on the test data are presented in Table 2. For English, HeidelTime's TempEval-3 tuned version achieves the best results, and all three HeidelTime versions outperform the systems of the eight other participating teams with a total number of 21 submissions (task A ranking measure *value F1*). For comparison, the results of the next best system (NavyTime) is listed in Table 2(a). For Spanish, we highly outperform the other two systems, as shown in Table 2(b).

Discussion. In order to be able to interpret HeidelTime's results on the training and test data, we performed an error analysis (TimeBank and Spanish training corpus). The most important findings are:

(i) For a rule-based system, HeidelTime's recall is relatively low (many false negatives; FN). However, note that several FN are intentional. 55% and 29% of 117 and 149 FN in the English and Spanish training corpora are due to imprecise expressions (*some time*; *the latest period*). These are difficult to normalize correctly, e.g., *some time* can refer to seconds or years. To guarantee high quality normalization, we do not extract expressions that cannot be normalized correctly with high probability.

(ii) There is a trade-off between precision and recall due to expressions referring to past, present, or future (X_REF). These are annotated either only in some contexts or inconsistently throughout the training data, and thus result in FN (21%/en; 34%/es) and false positives (43% of 98 FP in English training and 43%/es of 35 FP in Spanish training corpora).

(iii) The main sources for incorrect value normalization of underspecified expressions (*Feb. 1*; *Monday*) are wrongly detected reference times or relations to them (e.g., due to wrong tense identification), annotation errors in the corpora (e.g., *last week* annotated as WXX instead of the week it is referring to), granularity errors (e.g., *a year ago* can refer to a day, month, quarter, or year), and ambiguities (e.g., *the year* can be a duration or a specific year).

(iv) Some expressions in the Spanish test set were extracted and normalized correctly although no similar expressions exist in the Spanish training data. Here, the Spanish resources highly benefited from the high quality English resources as starting point of the development process, and from HeidelTime's language-independent normalization strategies.

(v) A reoccurring error in the English test set is that HeidelTime matches and normalizes expressions such as *two days earlier* while only *two days* should be annotated according to TimeML. This results in a relaxed match with false type and value.

5 Conclusions & Ongoing Work

In this paper, we presented HeidelTime's results in the TempEval-3 temporal tagging task. For both languages, English and Spanish, we achieved the best results of all participants (value F1). We showed that adding a new language to HeidelTime can result in high quality temporal tagging of the new language.

Currently, we are working on improving the Spanish tense detection to better normalize underspecified temporal expressions. Furthermore, we will make available HeidelTime resources for Arabic, Italian, and Vietnamese (HeidelTime, 2013).

References

Omar Alonso, Jannik Strötgen, Ricardo Baeza-Yates, and Michael Gertz. 2011. Temporal Information Retrieval: Challenges and Opportunities. In *Proceedings of the 1st International Temporal Web Analytics Workshop (TWAW 2011)*, pages 1–8.

HeidelTime. 2013. http://code.google.com/p/heideltime/.

Hector Llorens, Naushad UzZaman, and James F. Allen. 2012. Merging Temporal Annotations. In *19th International Symposium on Temporal Representation and Reasoning, TIME 2012*, pages 107–113.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, pages 321–324.

Jannik Strötgen and Michael Gertz. 2012a. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, Online first.

Jannik Strötgen and Michael Gertz. 2012b. Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3746–3753.

Jannik Strötgen, Michael Gertz, and Conny Junghans. 2011. An Event-centric Model for Multilingual Document Similarity. In *Proceeding of the 34rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11)*, pages 953–962.

Naushad UzZaman, Hector Llorens, James F. Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. *CoRR*, abs/1206.5333.

Marc Verhagen, Roser Sauri, 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.