WikiWarsDE: A German Corpus of Narratives Annotated with Temporal Expressions

Jannik Strötgen, Michael Gertz
Institute of Computer Science, Heidelberg University
Im Neuenheimer Feld 348, 69120 Heidelberg, Germany
E-mail: stroetgen@uni-hd.de, gertz@uni-hd.de

Abstract

Temporal information plays an important role in many natural language processing and understanding tasks. Therefore, the extraction and normalization of temporal expressions from documents are crucial preprocessing steps in these research areas, and several temporal taggers have been developed in the past. The quality of such temporal taggers is usually evaluated using annotated corpora as gold standards. However, existing annotated corpora only contain documents of the news domain, i.e., short documents with only few temporal expressions. A remarkable exception is the recently published corpus WikiWars, which is the first temporal annotated English corpus containing long narratives that are rich in temporal expressions. Following this example, in this paper, we describe the development and the characteristics of WikiWarsDE, a new temporal annotated corpus for German. Additionally, we present evaluation results of our temporal tagger HeidelTime on WikiWarsDE and compare them with results achieved on other corpora. Both, WikiWarsDE as well as our temporal tagger HeidelTime are publicly available.

Keywords: temporal expression, TIMEX2, corpus annotation, temporal information extraction

1. Introduction and Related Work

In the last decades, the extraction and normalization of temporal expressions have become hot topics in computational linguistics. In many research areas, temporal information plays an important role, e.g., in information extraction, document summarization, and question answering (Mani et al., 2005). In addition, temporal information is valuable in information retrieval and can be used to improve search and exploration tasks (Alonso et al., 2011). However, the tasks of extracting and normalizing temporal expressions are challenging due to the fact that there are many different ways to express temporal information in documents and that temporal expressions may be ambiguous.

Besides explicit expressions (e.g., “April 10, 2005”) that can directly be normalized to some standard format, relative and underspecified expressions are very common in many types of documents. To determine the semantics of such expressions, context information is required. For example, to normalize the expression “Monday” in phrases like “on Monday”, a reference time and the relation to the reference time have to be identified. Depending on the domain of the documents that are to be processed, this reference time can either be the document creation time or another temporal expression in the document. While the document creation time plays an important role in news documents, it is almost irrelevant in narrative style documents, e.g., documents about history or biographies. Despite these challenges, all applications using temporal information mentioned in documents rely on high quality temporal taggers, which correctly extract and normalize temporal expressions from documents.

Due to the importance of temporal tagging, there have been significant efforts in the area of temporal annotation of text documents. Annotation standards such as TIDES TIMEX2 (Ferro et al., 2005) and TimeML (Pustejovsky et al., 2003b; Pustejovsky et al., 2005) were defined and temporal annotated corpora like TimeBank (Pustejovsky et al., 2003a) were developed – although most of the corpora contain English documents only. Furthermore, research challenges were organized where temporal taggers were evaluated. The ACE

(Automatic Content Extraction) time expression and normalization (TERN) challenges were organized in 2004, 2005, and 2007.[1] In 2010, temporal tagging was one task in the TempEval-2 challenge (Verhagen et al., 2010). However, so far, research was limited to the news domain, i.e., the documents of the annotated corpora are short with only a few temporal expressions. The temporal discourse structure is thus usually easy to follow. Only recently, a first corpus containing narratives was developed (Mazur & Dale, 2010). This corpus, called WikiWars, consists of Wikipedia articles about famous wars in history. The documents are much longer than news documents and contain many temporal expressions. As the developers point out, normalizing the temporal expressions in such documents is more challenging due to the rich temporal discourse structure of the documents.

Motivated by this observation and by the fact that no temporal annotated corpus for German was publicly available so far, we created the WikiWarsDE corpus[2], which we present in this paper. WikiWarsDE contains the corresponding German articles of the documents of the English WikiWars corpus. For the annotation process, we followed the suggestions of the WikiWars developers, i.e., annotated the temporal expressions according to the TIDES TIMEX2 annotation standard using the annotation tool Callisto[3]. To be able to use publicly available evaluation scripts, the format of the ACE TERN corpus was selected. Thus, evaluating a temporal tagger on the WikiWarsDE corpus is straightforward and evaluation results of different taggers can be compared easily.

The remainder of the paper is structured as follows. In Section 2, we describe the annotation schema and the corpus creation process. Then, in Section 3, we present detailed information about the corpus such as statistics on the length of the documents and the number of temporal expressions. In addition, evaluation results of our own temporal tagger on the WikiWarsDE corpus are presented. Finally, we conclude our paper in Section 4.

[1] The 2004 and 2005 training sets and the 2004 evaluation set are released by the LDC as is the TimeBank corpus; see http://www.ldc.upenn.edu/

[2] WikiWarsDE is publicly available on http://dbs.ifi.uni-heidelberg.de/temporal_tagging/

[3] http://callisto.mitre.org/

Temporal Expression	Value of the VAL attribute
November 12, 2001	2001-11-12
9:30 p.m.	2001-11-12T21:30[4]
24 months	P20M
daily	XXXX-XX-XX

Table 1: Normalization examples (VAL) of temporal expressions of the types date, time, duration, and set.

2. Annotation Schema and Corpus Creation

In Section 2.1, we describe the annotation schema, which we used for the annotation of temporal expressions in our newly created corpus. Furthermore, we explain the task of normalizing temporal expressions using some examples. Then, in Section 2.2, we detail the corpus creation process and explain the format, in which WikiWarsDE is publicly available.

2.1. Annotation Schema

Following the approach of Mazur and Dale (2010), we use TIDES TIMEX2 as annotation schema to annotate the temporal expressions in our corpus. The TIDES TIMEX2 annotation guidelines (Ferro et al., 2005) describe how to determine the extents of temporal expressions and their normalizations. In addition to date and time expressions, such as “November 12, 2001” and “9:30 p.m.”, temporal expressions describing durations and sets are to be annotated as well. Examples for expressions of the types duration and set are “24 months” and “daily”, respectively.

The normalization of temporal expressions is based on the ISO 8601 standard for temporal information with some extensions. The following five features can be used to normalize a temporal expression:

- VAL (value)
- MOD (modifier)
- ANCHOR_VAL (anchor value)
- ANCHOR_DIR (anchor direction)
- SET

The most important feature of a TIMEX2 annotation is the “VAL” (value) feature. For the four examples above, the values of VAL are given in Table 1. Furthermore, “MOD” (modifier) is used, for instance, for expressions

[4] Assuming that “9:30 p.m.” refers to 9:30 p.m. on November 12, 2001.

such as “the end of November 2001”, where MOD is set to “END”, i.e., to capture additional specifications not captured by VAL. ANCHOR_VAL and ANCHOR_DIR are used to anchor a duration to a specific date, using the value information of the date and specifying whether the duration starts or ends on this date. Finally, SET is used to identify set expressions.

Often, for example in the TempEval-2 challenge, the normalization quality of temporal taggers is evaluated based on the VAL (value) feature, only. This fact points out the importance of this feature and was the motivation to evaluate the normalization quality of our temporal tagger based on this feature as described in Section 3.

2.2. Corpus Creation

For the creation of the corpus, we followed Mazur and Dale (2010), the developers of the English WikiWars corpus. We selected the 22 corresponding German Wikipedia articles and manually copied sections describing the course of the wars.[5] All pictures, cross-page references, and citations were removed. All text files were then converted into SGML files, the format of the ACE TERN corpora containing “DOC”, “DOCID”, “DOCTYPE”, “DATETIME”, and “TEXT” tags. The document creation time was set to the time of downloading the articles from Wikipedia. The “TEXT” tag surrounds the text that is to be annotated.

Similar to Mazur and Dale (2010), we used our own temporal tagger, which is described in Section 3.2, containing a rule set for German as a first-pass annotation tool. The output of the tagger can then be imported to the annotation tool Callisto for manual correction of the annotations. Although this fact has to be taken into account when comparing the evaluation results on WikiWarsDE of our temporal tagger with other taggers, this procedure is motivated by the fact that “annotator blindness” is reduced to a minimum, i.e., that annotators miss temporal expressions. Furthermore, the annotation effort is reduced significantly since one does not have to create a TIMEX2 tag for the expressions already identified by the tagger.

At the second annotation stage, the documents were examined for temporal expressions missed by the temporal tagger and annotations created by the tagger were manually corrected. This task was performed by two annotators – although Annotator 2 only annotated the extents of temporal expressions. The more difficult task of normalizing the temporal expressions was performed by Annotator 1 only, since a lot of experience in temporal annotation is required for this task. At the third annotation stage, the results of both annotators were merged and in cases of disagreement the extents and normalizations were rechecked and corrected by Annotator 1.

To compare our inter-annotator agreement for the determination of the extents of temporal expressions to others, we calculated the same measures as the developers of the TimeBank-1.2 corpus. They calculated the average of precision and recall with one annotator's data as the key and the other's as the response. Using a subset of ten documents, they report inter-annotator agreement of 96% and 83% for partial match (lenient) and exact match (strict), respectively.[6] Our scores for lenient and exact match on the whole corpus are 96.7% and 81.3%, respectively.

Finally, the annotated files, which contain inline annotations, were transformed into the ACE APF XML format, a stand-off markup format used by the ACE evaluations. Thus, the WikiWarsDE corpus is available in the same two formats as the WikiWars corpus, and the evaluation tools of the ACE TERN evaluations can be used with this German corpus as well.

3. Corpus Statistics and Evaluation Results

In this section, we first present some statistical information about the WikiWarsDE corpus, such as the length of the documents and the number of temporal expressions in the documents (Section 3.1). Then, in Section 3.2, we shortly introduce our own temporal tagger HeidelTime, present its evaluation results on WikiWarsDE, and compare them with results achieved on other corpora.

[5] Due to the shortness of the Wikipedia article about the Punic Wars in general, we used sections of three separate articles about the 1st, 2nd, and 3rd Punic Wars.

[6] For more information on TimeBank, see http://timeml.org/site/timebank/documentation-1.2.html.

Corpus	Docs	Token	Timex	Token / Timex	Timex / Document
ACE 04 en train	863	306,463	8,938	34.3	10.4
TimeBank 1.2	183	78,444	1,414	55.5	7.7
TempEval2 en train	162	53,450	1,052	50.8	6.5
TempEval2 en eval	9	4,849	81	59.9	9.0
WikiWars	22	119,468	2,671	44.7	121.4
WikiWarsDE	22	95,604	2,240	42.7	101.8

Table 2: Statistics of the WikiWarsDE corpus and other publicly available or released corpora.

3.1. Corpus Statistics

The WikiWarsDE corpus contains 22 documents with a total of more than 95,000 tokens and 2,240 temporal expressions. Note that the fact that the WikiWars corpus contains almost 25,000 tokens more than WikiWarsDE can be partly explained by the differences between the two languages. In German compounds are very frequent, e.g., the 3 English tokens "course of war" is just 1 token in German (“Kriegsverlauf”).

In Table 2, we present some statistics of the corpus in comparison to other publicly available corpora. On the one hand, the density of temporal expressions (Token/Timex) is similar among the documents of all the corpora. In WikiWarsDE, one temporal expression occurs every 42.7 tokens on average.

On the other hand, one can easily see that the documents of the WikiWarsDE and the WikiWars corpora are much longer and contain many more temporal expressions than the documents of the news corpora. While WikiWars and WikiWarsDE contain 121.4 and 101.8 temporal expressions per document on average, the number of temporal expressions on the news copora ranges between 6.5 and 10.4 temporal expressions only. Thus, the temporal discourse structure is much more complex for the narrative-style documents in WikiWars and WikiWarsDE. Further statistics on the single documents of WikiWarsDE are published with the corpus.

3.2. Evaluation Results

After the development of the corpus, we evaluated our temporal tagger HeidelTime on the corpus. HeidelTime is a multilingual, rule-based temporal tagger. Currently, two languages are supported (English and German), but due to the strict separation between the source code and the resources (rules, extraction patterns, normalization information), HeidelTime can be easily adapted to further languages. In the TempEval-2 challenge, HeidelTime achieved the best results for the extraction and normalization of temporal expressions from English documents (Strötgen & Gertz, 2010; Verhagen et al., 2010). Since HeidelTime uses different normalization strategies depending on the type of the documents that are to be processed (news- or narrative-style documents), we were able to show that HeidelTime achieves high quality results on both kinds of documents for English.[7]

With the development of WikiWarsDE, we are now able to evaluate HeidelTime on a German corpus as well. For this, we use the well-known evaluation measures of precision, recall, and f-score. In addition, we distinguish between lenient (overlapping match) and strict (exact match) measures. For the normalization, one can calculate the measures for all expressions that were correctly extracted by the system (value). This approach is used by the ACE TERN evaluations. However, similar to Ahn et al. (2005) and Mazur and Dale (2010), we argue that it is more meaningful to combine the extraction with the normalization tasks, i.e., to calculate the measures for all expressions in the corpus (lenient+value and strict+value).

[7] More information on HeidelTime, its evaluation results on several corpora, as well as download links and an online demo can be found at http://dbs.ifi.uni-heidelberg.de/heideltime/.

Corpus	lenient			strict			value			lenient + value			strict + value		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
TimeBank 1.2	90.5	91.4	90.9	83.5	84.3	83.9	86.2	86.2	86.2	78.0	78.8	78.4	73.2	74.0	73.6
WikiWars	93.9	82.4	87.8	86.0	75.4	80.4	89.5	90.1	89.8	84.1	73.8	78.6	79.6	69.8	74.4
WikiWarsDE	98.5	85.0	91.3	92.6	79.9	85.8	87.0	87.0	87.0	85.7	74.0	79.4	82.5	71.2	76.5

Table 3: Evaluation results of our temporal tagger on an English news corpus (TimeBank 1.2), an English narratives corpus (WikiWars) and our newly created German narratives corpus WikiWarsDE.

On WikiWarsDE, HeidelTime achieves f-scores of 91.3 and 85.8 for the extraction (lenient and strict, respectively) and 79.4 and 76.5 for the normalization (lenient + value and strict + value, respectively).

For comparison, we present the results of HeidelTime on some English corpora. As shown in Table 3, our temporal tagger achieves equally good results on both, the narratives corpora (WikiWars and WikiWarsDE) and the news corpus (TimeBank). Note that our temporal tagger uses different normalization strategies depending on the type of the corpus that is to be processed. This might be the main reason why HeidelTime clearly outperforms the temporal tagger of the WikiWars developers. For the WikiWars corpus, Mazur and Dale (2010) report f-scores for the normalization of only 59,0 and 58,0 (lenient + value and strict + value, respectively). Compared to these values, HeidelTime achieves much higher f-scores, namely 78.6 and 74.4, respectively.

4. Conclusions

In this paper, we described WikiWarsDE, a temporal annotated corpus containing German narrative-style documents. After presenting the creation process and statistics of WikiWarsDE, we used the corpus to evaluate our temporal tagger HeidelTime. While Mazur and Dale (2010) report lower evaluation results of their temporal tagger on narratives than on news documents, HeidelTime achieves similar results on both types of corpora. Nevertheless, we share their opinion that the normalization of temporal expressions on narratives is challenging. However, using different normalization strategy for different types of documents (news-style and narrative-style documents), this problem can be tackled.

By making available WikiWarsDE and HeidelTime, we provide useful contributions to the community in support of developing and evaluating temporal taggers and of improving temporal information extraction.

5. Acknowledgements

We thank the anonymous reviewers for their valuable suggestions to improve the paper.

6. References

Ahn, D., Adafre, S.F., de Rijke, M. (2005): Towards Task-Based Temporal Extraction and Recognition. In G. Katz, J. Pustejovsky, F. Schilder (Eds.), Extracting and Reasoning about Time and Events. Dagstuhl, Germany: Dagstuhl Seminar Proceedings.

Alonso, O., Strötgen, J., Baeza-Yates, R., Gertz, M. (2011): Temporal Information Retrieval: Challenges and Opportunities. In Proceedings of the 1st International Temporal Web Analytics Workshop (TWAW), pp. 1–8.

Ferro, L., Gerber, L., Mani, I., Sundheim, B., Wilson, G. (2005): TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report, The MITRE Corporation.

Mani, I., Pustejovsky, J., Gaizauskas, R.J. (2005): The Language of Time: A Reader. Oxford University Press.

Mazur, P., Dale, R. (2010): WikiWars: A New Corpus for Research on Temporal Expressions. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 913-922.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R.J., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., Lazo, M. (2003a): The

TIMEBANK Corpus. In Proceedings of Corpus Linguistics 2003, pp. 647–656.

Pustejovsky, J., Castaño, J.M., Ingria, R., Sauri, R., Gaizauskas, R.J., Setzer, A., Katz, G. (2003b): TimeML: Robust Specification of Event and Temporal Expressions in Text. In New Directions in Question Answering, pp 28–34.

Pustejovsky, J., Knippen, R., Littman, J., Sauri, R. (2005): Temporal and Event Information in Natural Language Text. Language Resources and Evaluation, 39(2-3):123–164.

Strötgen, J., Gertz, M. (2010): HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval), pp. 321–324.

Verhagen, M., Sauri, R., Caselli, T., Pustejovsky, J. (2010): SemEval-2010 Task 13: TempEval-2. In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval), pp. 57–62.