Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards

Jannik Strötgen, Michael Gertz

Institute of Computer Science, Heidelberg University
Im Neuenheimer Feld 348, 69120 Heidelberg, Germany
{stroetgen, gertz}@uni-hd.de

Abstract

In the last years, temporal tagging has received increasing attention in the area of natural language processing. However, most of the research so far concentrated on processing news documents. Only recently, two temporal annotated corpora of narrative-style documents were developed, and it was shown that a domain shift results in significant challenges for temporal tagging. Thus, a temporal tagger should be aware of the domain associated with documents that are to be processed and apply domain-specific strategies for extracting and normalizing temporal expressions. In this paper, we analyze the characteristics of temporal expressions in different domains. In addition to news- and narrative-style documents, we add two further document types, namely colloquial and scientific documents. After discussing the challenges of temporal tagging on the different domains, we describe some strategies to tackle these challenges and describe their integration into our publicly available temporal tagger HeidelTime. Our cross-domain evaluation validates the benefits of domain-sensitive temporal tagging. Furthermore, we make available two new temporally annotated corpora and a new version of HeidelTime, which now distinguishes between four document domain types.

Keywords: temporal tagging, corpora comparison, HeidelTime

1. Introduction

Temporal information occurs in many types of text documents, and its extraction and normalization are crucial for many natural language processing and understanding tasks. Thus, temporal information extraction is an active research field and lots of efforts have been undertaken in the past to develop temporal taggers such as Chronos (Negri and Marseglia, 2004), DANTE (Mazur and Dale, 2007), and HeidelTime (Strötgen and Gertz, 2010), to create temporal annotated corpora, e.g., the TimeBank corpus (Pustejovsky et al., 2003), and to organize challenges on temporal information extraction like the ACE TERN and TempEval competitions[1]. So far, most of the work on temporal information extraction concentrated on documents from the news domain, resulting in a couple of temporal annotated news corpora and temporal taggers that achieve good results on documents from this domain.

Recently, two temporal annotated corpora, WikiWars (Mazur and Dale, 2010) and its German counterpart WikiWarsDE (Strötgen and Gertz, 2011), were developed, which contain narrative-style documents. Due to the domain shift, there are new challenges for temporal taggers to extract and normalize temporal expressions from these documents. As Mazur and Dale (2010) showed, switching between domains results in a significant performance loss of temporal tagging, for both extraction and normalization. However, Kolomiyets et al. (2011) studied the model portability of a machine learning based temporal tagger and showed on a small set of manually annotated Wikipedia documents that the extraction performance drop can be avoided by synonym-based bootstrapping. In addition, we showed that temporal taggers can achieve high quality results for the extraction and the normalization of temporal expressions on news and narrative-style documents, if applying different normalization strategies depending on the domain of the documents that are to be processed (Strötgen and Gertz, 2011).

Motivated by this observation, in this paper, we compare temporal tagging on news and narrative documents with temporal tagging on two further domains, namely texts from colloquial (SMS) and scientific (biomedical) documents. By performing a detailed analysis, we show that (i) the types of temporal expressions occurring in different domains vary, (ii) the challenges for temporal taggers are domain-dependent, and (iii) these challenges can be tackled by providing information of the document type to the tagger, which may then apply different tagging strategies depending on the domain. Furthermore, by making available the newly annotated colloquial and scientific corpora and a new version of our temporal tagger HeidelTime, we provide valuable contributions to the community.[2]

The remainder of the paper is structured as follows. After surveying existing annotation standards and temporal annotated corpora, in Section 2, we describe what types of temporal expressions exist and how they can occur in text documents in Section 3. In Section 4, we present the corpus creation and annotation process, before analyzing the challenges of temporal tagging on different domains as well as strategies to address them in Section 5. Finally, we demonstrate the need for a temporal tagger to be domain-sensitive by performing a cross-domain evaluation in Section 6.

[1] The ACE TERN and TimeBank corpora are released by the Linguistic Data Consortium, see: http://www.ldc.upenn.edu/. For details on TempEval, see: http://www.timeml.org/.

[2] HeidelTime, WikiWarsDE, and the newly annotated corpora (Time4SMS and Time4SCI) are publicly available at http://dbs.ifi.uni-heidelberg.de/heideltime/. The new version of HeidelTime distinguishes between colloquial and scientific text in addition to news- and narrative-style documents.

2. Annotation Standards and Annotated Corpora

There are two standards for annotating temporal information in documents: TIDES TIMEX2 (Ferro et al., 2005) and TimeML (Pustejovsky et al., 2005). Both standards present guidelines on how to determine the extents and how to normalize the values of temporal expressions. TimeML's TIMEX3 tags, which are based on TIMEX2, are used to annotate temporal expressions. In addition, annotations for several other temporal phenomena such as temporal relations between events are defined in TimeML.

Although the differences between the TIMEX2 and TIMEX3 guidelines are significant, it is possible to automatically translate annotations from one standard into the other, as was recently shown by Saquete and Pustejovsky (2011). In addition, the features of both tags are quite similar to each other. For this, we only describe the features of TIMEX2 tags, which we use for the newly annotated corpora as described in Section 4. According to the TIDES TIMEX2 annotation guidelines, a TIMEX2 tag may contain the following attributes:

- VAL: a normalized form of the expressions based on the ISO 8601 standard for temporal information
- MOD: temporal modifiers
- ANCHOR_VAL: a normalized form of an anchoring date/time
- ANCHOR_DIR: the relative direction between VAL and ANCHOR_VAL
- SET: to identify expressions denoting sets of times

Using these attributes, the four types of temporal expressions (dates, times, durations, and sets) can be normalized. The value attribute (VAL) of date and time expressions directly refers to a point in time, e.g., "2011-02-15" for the expression "February 15, 2011". For durations and set expressions, it covers the length of the time interval, e.g., "P2D" for "two days" and "every two days". To distinguish durations and sets, the SET attribute is set to true for all set expressions. Finally, durations may be anchored to some point in time using the ANCHOR_VAL and ANCHOR_DIR attributes to refer to a normalized value of a time or date expression and to define the temporal relation to the anchor, respectively.

Based on the two annotation standards, several temporal annotated corpora were developed in the past. The ACE TERN (time expression and normalization) contests 2004, 2005, and 2007 used TIMEX2 annotation guidelines to develop the training and annotation corpora.[3] As a reference corpus for TimeML, the TimeBank corpus (Pustejovsky et al., 2003) was developed during the Time and Event Recognition for Question Answering workshop (TERQAS). In addition to temporal expressions (TIMEX3 tags), events and temporal relations are annotated. This corpus was used as a training corpus for the TempEval-2 challenge (Verhagen et al., 2010) while the evaluation corpus was created using newly annotated documents. All these corpora[4] contain news documents. In addition, the ACE TERN 2005 and 2007 corpora contain conversations, discussions, and webblogs, which can also be regarded as news-style documents since the document creation time is important to normalize temporal expressions in the same way as for news documents. In general, the documents of these corpora are short and contain only a few temporal expressions.

Due to the lack of temporal annotated corpora of other domains, WikiWars and its German counterpart WikiWarsDE were developed recently. These contain descriptions of famous wars in history, i.e., narrative style documents, which are much longer than the news-style documents of the other corpora. Furthermore, as we will detail in Section 4, they contain many more temporal expressions and thus a more complex discourse structure. In summary, so far, there have been temporal annotated corpora containing news- and narrative-style documents. However, to the best of our knowledge, there is no research on temporal tagging on further domains, as it is discussed in this paper.

3. Temporal Expressions in Documents

Following TimeML, the standard mark-up language for temporal annotation, there are four types of temporal expressions: dates (May 23, 2012), times (3 p.m.), durations (three weeks), and sets (daily). Furthermore, there are different ways of how expressions of the types date and time can be realized in text, namely either explicitly, implicitly, relative, or underspecified. Since temporal expression can be normalized, as indicated in the previous section, a standard value can be associated with each temporal expression. However, depending on how an expression occurs in a document, the normalization may be challenging. While explicit expressions (e.g., May 23, 2012) can be normalized directly, additional knowledge is needed to normalize expressions of the other types. To normalize implicit expressions such as holidays (e.g., Columbus Day 2010), knowledge about their meaning is required. Relative expressions (e.g., two days later) require the identification of the reference time. To normalize underspecified expressions (e.g., November) additionally the temporal relation to the identified reference time is needed.

One of the main challenges for the normalization of temporal expressions is to identify the correct reference time and the temporal relation to this reference time. As shown in previous work (Strötgen and Gertz, 2011), it is useful that a temporal tagger is aware of the domain of the documents that are to be processed, i.e., that a temporal tagger applies different normalization strategies depending of the domain of the documents.

More details on the challenges of the extraction and normalization of temporal expressions on news- and narrative-style documents, but also on colloquial and scientific text, are given in Section 5.

[3] See, http://www.itl.nist.gov/iad/mig/tests/ace/.

[4] The TimeBank corpus, the ACE TERN 2004 training and evaluation, and the 2005 training corpora are released by the Linguistic Data Consortium (http://www.ldc.upenn.edu/). The TempEval-2 corpora are publicly available on http://timeml.org/site/timebank/timebank.html.

4. Annotating Colloquial and Scientific Text

Due to different challenges for temporal tagging of either news or narrative-style documents, in this paper, we study the challenges arising in further domains, namely in colloquial (SMS) and scientific (biomedical) texts. In both types of documents, temporal information plays a crucial role, e.g., in SMS messages for communicating about upcoming events or meetings, and in biomedical documents – as representative of scientific texts – for describing chronological procedures such as clinical trials. Since there were no temporal annotated corpora with documents from these domains available so far, we created two new corpora and manually annotated the temporal expressions occurring in the documents, as described next.

4.1. Corpus Creation

In Section 4.1.1 and Section 4.1.2, we describe the document selection for the colloquial and the biomedical corpus, respectively.

4.1.1. Colloquial Corpus Creation

Although there are some SMS corpora publicly available, there are four main requirements for the SMS corpus to be applicable for publishing a temporal annotated SMS corpus: (i) it has to be freely available to allow others to reproduce the corpus, (ii) the language of the messages has to be English since when developing a corpus for a new domain, we believe that English annotated corpora are most valuable for the research community, (iii) the corpus has to be large since the single messages are short and thus cannot contain many temporal expressions, and (iv) the document creation time (i.e., the time when the message was sent) has to be available for the messages.

The availability of the sending time is crucial for normalizing underspecified and relative temporal expressions, which we expect to occur frequently in SMS texts. Due to these requirements, we used the NUS SMS corpus (Chen and Kan, 2011) as basis of our colloquial corpus. However, the 2004 version of the corpus does not fulfill all our requirements, since these documents do not contain information about the sending time. Without the documents of the 2004 version, the corpus contains 28,268 messages (June 2011 version)[5]. Due to privacy reasons, the developers of the corpus anonymize all SMS automatically and sensitive data are substituted by placeholders. Unfortunately, multi-digit numbers and some specific time information are part of this sensitive data. To overcome this problem, we replaced these placeholders of digits and times by some standard values in the original format.[6] Then, we randomly selected 1,000 documents as our SMS corpus called Time4SMS, in which we manually annotated all occurring temporal expressions as described in Section 4.2.

4.1.2. Scientific Corpus Creation

As the second domain for our temporal analysis we chose scientific documents. However, temporal expressions are only frequent in some kinds of scientific literature. A good representative of scientific documents containing many temporal expressions are texts from the biomedical domain, for example, publications describing clinical trials. For selecting documents, we used PubMed[7], which contains citations with abstracts and metadata such as publication dates of more than 20 million publications of the biological and biomedical domain. Using the PubMed search interface, we queried for "clinical trials" and downloaded the abstracts and metadata of the 50 most recent publications as our scientific corpus called Time4SCI.

In the next section, we describe how the documents were formatted and annotated with temporal expressions. Furthermore, in Section 4.3 we detail the characteristics of the corpora, e.g., the length of the documents and the number of temporal expressions in the documents, and compare them with other publicly available corpora.

Corpus	Doc	Token	Timex	Token/ Doc	Timex/ Doc
TimeBank	183	78,444	1,414	428.7	7.7
WikiWars	22	119,468	2,671	5430.4	121.4
Time4SMS	1,000	20,176	1,341	20.2	1.3
Time4SCI	50	19.194	317	383.9	6.3

Table 1: Statistics of temporal annotated corpora.

4.2. Annotation Procedure

As for the annotation of WikiWarsDE (Strötgen and Gertz, 2011), we followed the developers of WikiWars (Mazur and Dale, 2010), i.e., we formatted the corpora in SGML, the format of the ACE TERN corpora. This makes it possible to evaluate temporal taggers on our newly annotated corpora using the publicly available TERN evaluation scripts[8]. For the annotation of temporal expressions, we used the TIDES TIMEX2 format (Ferro et al., 2005) with its attributes described in Section 2. We performed a three phase annotation process: (i) automatic pre-annotation, (ii) manual annotation with correcting wrong and adding missing expressions, (iii) manual merging and validation of the annotations. The evaluation process was carried out in the same way as for WikiWarsDE. For details on the annotation procedure, we refer to (Strötgen and Gertz, 2011).

During the manual annotation process, we were faced with domain-specific difficulties. Due to many unresolvable temporal expressions in the scientific corpus, we suggest a new way to normalize these expressions. However, since the normalization of unresolvable expressions is one of the main challenges of temporal tagging scientific documents, the details of this issue and how it can be addressed are described in Section 5.1 and Section 5.2, respectively.

In contrast to news and narrative-style Wikipedia documents, it is very challenging to annotate colloquial and scientific text since deep domain knowledge is needed to fully understand such documents. For this, we regard our newly developed annotated corpora as preliminary versions of a gold standard.

[5] http://wing.comp.nus.edu.sg/SMSCorpus/

[6] The NUS SMS corpus developers kindly provided their function to replace sensitive data, so that we were able to reproduce standard values for the placeholders in the original format.

[7] http://www.ncbi.nlm.nih.gov/pubmed/

[8] http://fofoca.mitre.org/tern.html

4.3. Corpora Statistics

In Table 1, we compare our newly annotated corpora to a news corpus (TimeBank) and a narrative-style corpus (WikiWars) with respect to the number of documents, tokens, and temporal expressions.

The documents of the Time4SMS corpus are very short while the Time4SCI documents are similar to the news documents with respect to the average document length. Due to their shortness, documents in the SMS corpus contain only few temporal expressions. The average number of temporal expressions in the clinical-trial documents and in the news documents is comparable. In contrast, the narrative documents are very long and contain many temporal expressions. Although the Time4SMS and Time4SCI corpora are smaller than TimeBank and WikiWars with respect to the number of tokens, their sizes are sufficient to discover significant differences between the corpora resulting in different challenges for temporal tagging. These challenges and strategies to address them are detailed next.

5. Temporal Tagging on Different Domains

While temporal tagging on news documents was the focus of research in the past, temporal tagging on narrative-style Wikipedia documents was studied recently as well. However, there is only very little work on temporal tagging on colloquial or scientific text and on comparing temporal tagging on different domains. In this section, we describe the challenges that occur when processing different domains (Section 5.1) and strategies how the challenges can be addressed (Section 5.2) by comparing the characteristics of our newly annotated corpora (Time4SMS, Time4SCI) with a news corpus (TimeBank) and a narrative corpus (WikiWars).

5.1. Challenges on Different Domains

To identify the different challenges for temporal tagging on different domains, we analyze the temporal expressions occurring in the four corpora. The number of document creation times (DCTs) in a corpus equals the number of documents and thus, the percentage of DCTs in corpora containing long documents with many temporal expressions is very low (WikiWars), but very high for corpora with short documents (Time4SMS). Since the DCT is usually easy to extract and normalize, we concentrate on the other types in our further analysis. In Figure 1, the frequencies of the different types of temporal expressions are shown. This directly results in the first challenge for temporal tagging on different domains:

Challenge 1: Broad Coverage
There is a need that all four types of temporal expressions (dates, times, durations, and sets) are well covered by a temporal tagger.

In all corpora, expressions of the type date are frequent. In contrast, time expressions are only frequent in the SMS corpus and set expressions are well covered only in the clinical-trial corpus. Duration expressions occur in all corpora, however, they are most frequent in the clinical-trial corpus. Thus, when developing a temporal tagger on one domain only, e.g., on the news domain (as are most existing systems), this may result in a worse coverage on the other domains since not all types of expressions may be covered very well. For example, it would be possible to extract more than 80% of the temporal expressions from the news and the narrative corpora with a temporal tagger that only extracts date expressions. In summary, a broad coverage of a temporal tagger is less important when processing domains, in which mostly one type of temporal expressions occurs (i.e., news and narratives), while it is more important when processing domains such as SMS and clinical-trial documents, in which there are many temporal expressions of different types.

For a deeper analysis of the corpora, we explore date and time expressions in more detail. These may be either explicit, implicit, relative, or underspecified resulting in different challenges for temporal tagging, and especially for the normalization of temporal expressions (see Section 3). Figure 2 shows the corresponding distributions for the four corpora. The simply normalizable explicit expressions are frequent in WikiWars (51.6%) while they hardly occur in Time4SMS (0.3%). Implicit expressions are rare in all the four corpora. However, to extract and normalize the occurring implicit expressions, the temporal tagger requires additional knowledge resources. For example, to extract and normalize holidays and expressions such as "D-Day", they have to be known by the tagger in the same way as usual temporal words such as names of months. Thus, the second challenge can be described as follows:

Challenge 2: Resources for Implicit Expressions
If the documents of a specific domain contain many implicit expressions, there is a need to easily add resources to extract and normalize them.

Although there are not many implicit expressions in the four corpora, the occurring implicit expressions can only be extracted if the temporal tagger can access resources containing information about them.

To normalize relative and underspecified expressions, e.g., "next Monday" or "November" in phrases such as "In November", the temporal tagger has to identify the reference time.

Challenge 3: Reference Time Identification
To be able to normalize relative and underspecified expressions, the temporal tagger has to identify the correct reference time.

In the news and SMS corpora, the identification of the reference time is relatively simple since it is the document creation time (DCT) in most cases. 78.1% (news) and 85.5% (SMS) of the time and date expressions are either relative or underspecified expressions with the DCT being the reference time while there are only 10% (news) and no (SMS) expressions for which the reference time is another temporal expression in the text itself. In narrative-style documents, almost always the reference time has to be identified in the documents' texts. To normalize 44.7% of the time and date expressions the reference time has to be identified in the documents' text while only 0.3% have the DCT as reference time. Furthermore, due to the large number

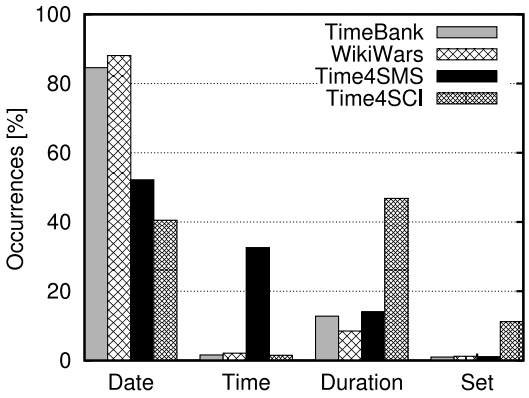


Figure 1: Types of temporal expressions in the analyzed corpora. Document creation times are excluded.

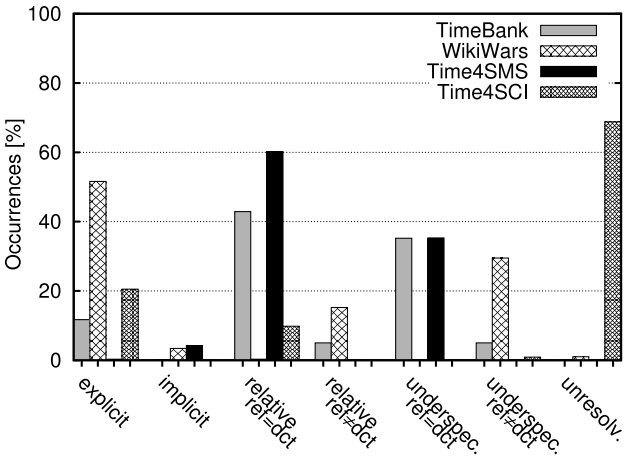


Figure 2: Characteristics of time and date expressions on the analyzed corpora (ref: reference time; dct: document creation time).

of temporal expressions in the documents of WikiWars, the temporal discourse structure is more complex, i.e., the reference time identification task is even more challenging in narrative-style documents. In the clinical-trial corpus relative and underspecified expressions are rare, but if they occur their reference time is usually the DCT.

In summary, it is more challenging to identify the reference time in narrative-style documents than in the other domains since it has to be identified in the text and is usually not the DCT. Thus, to address Challenge 3, a temporal tagger should apply different strategies depending on the domain to identify the reference time of relative and underspecified expressions as described in Section 5.2.

In contrast to normalizing relative expressions, for the normalization of underspecified expressions, it is not sufficient to identify the reference time, but the relation to the reference time is also needed.

Challenge 4: Identification of the Relation to the Reference Time
To normalize underspecified expressions the relation to the reference time has to be identified.

This is a challenging task on all domains. As detailed in the next section, if the DCT is the reference time, a good strategy will be to identify the tense of the sentence in which the expression occurs. If the tense cannot be identified, e.g., several SMS texts do not contain a verb at all, the normalization will be even more challenging and the relationship has to be guessed. While news often describe events that already happened, the analysis of the Time4SMS corpus suggests that SMS messages tend to refer to upcoming events. If the reference time is not the DCT, one may assume that there is a chronological order in the text, i.e., that the underspecified expressions refers to a point in time after a previously mentioned reference time. Thus, domain-dependent strategies to address Challenge 4 are needed in these cases, as will also be described in Section 5.2. While there are hardly any underspecified expressions in the clinical-trial corpus, addressing Challenge 4 is crucial for processing the news, narrative, and SMS corpora.

In SMS documents other kinds of challenges arise additionally, which do hardly occur in neither news, narrative, nor scientific documents. These can be summarized as follows:

Challenge 5: Coping with Non-standard Language
In some domains, non-standard language issues may occur frequently.

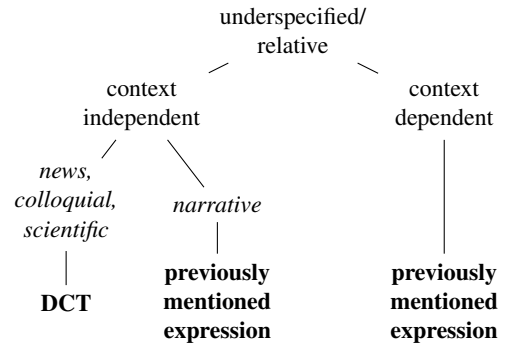
Challenge 5 can further be split up into the following issues:

- broad variety of spelling variations and word creations (e.g., “night”, “nite”, “nit”, “ni8”)
- type errors (e.g., “mornimg”)
- missing spaces (e.g., “todaygot . . . ”)

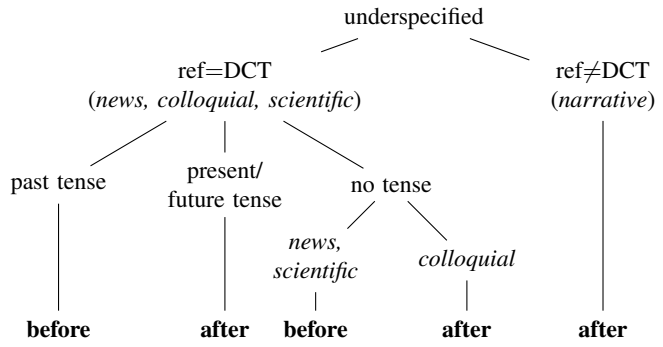
These issues usually occur only in colloquial documents and should thus be handled by a temporal tagger if colloquial text is processed. Thus, Challenge 5 is only relevant for the colloquial corpus while it does not occur when processing documents from the other corpora.

An additional challenge in the SMS corpus, as in every other corpus containing parts of conversations, is that required context information may have been mentioned in previous messages but cannot be accessed for the normalization. This challenge can only be addressed if the conversation (e.g., several SMS that build a conversation) is processed by the temporal tagger as a single document. Thus, this challenge is not a challenge that can be addressed by the temporal tagger itself, but may be addressed during corpus preprocessing.

While Challenge 5 is only relevant for the SMS corpus, there is another challenge that is mainly relevant for the clinical-trial corpus since it affects many temporal expressions in this corpus. Often, these documents contain their own time frame, e.g., the beginning of a clinical trial. This results in the fact, that expressions such as “on day 3” or “after three weeks” cannot be normalized to a real point in time. In the clinical-trial corpus, almost 70% of the time and date expressions refer to a local time frame. However, instead of normalizing such expressions to unspecific days (XXXX-XX-XX) as suggested in the annotation guidelines, we suggest to create a local time frame for every document. Thus, this challenge that will be further detailed in the next section can be formulated as follows:



(a) Identification of the reference time (bold) with domain-dependent strategies (italics).



(b) Identification of the temporal relation to the reference time (bold) with domain-dependent strategies (italics).

Figure 3: Strategies to identify the reference time (a) and the temporal relation to the reference time (b).

Challenge 6: Local Normalization of Unresolvable Expressions
Time and date expressions that cannot be normalized to a global point in time should be normalized with respect to a local time frame.

In summary, there are several challenges for temporal tagging. While some of them arise only when processing specific domains (e.g., challenges 5 and 6), others may occur independent of the domain, e.g., to identify the reference time of relative and underspecified expressions. However, due to the different characteristics of documents from different domains, it is necessary to tackle the challenges in a domain-dependent manner. In the next section, such domain-dependent strategies will be suggested.

5.2. Domain-dependent Strategies

In this section, general and domain-dependent strategies to address the challenges described in the previous section are suggested. Furthermore, we detail which of these strategies are used by our publicly available temporal tagger HeidelTime and how they are realized.

5.2.1. HeidelTime

HeidelTime is a rule-based, multilingual, cross-domain temporal tagger. Its architecture (Strötgen and Gertz, 2012) strictly separates between the source code and the language-dependent resources. For every language, HeidelTime contains resources such as patterns, normalization information, and rules, which can easily be modified or extended. While the first version of HeidelTime was built to process news and narratives, the current version also processes colloquial and scientific texts. To use HeidelTime, one has to specify the language and the domain of the documents that are to be processed.

5.2.2. Addressing Challenges 1 and 2

Challenge 1, i.e., that a temporal tagger should cover temporal expressions of all types (dates, times, durations, and sets) adequately, can be tackled if a temporal tagger is developed based on data of all domains that shall be processed. Either a machine learning based temporal tagger should be trained using training data of all domains or the rules of a rule-based temporal tagger should be developed based on examples of all domains.

The second challenge of a temporal tagger, i.e., that implicit temporal expressions can be integrated easily, can be solved if the architecture of a temporal tagger supports the simple integration of additional resources. While the vocabulary of many temporal expressions is limited, e.g., based on numbers and names of months and days, the vocabulary of implicit expressions is potentially unlimited. Thus, to extract and normalized these expressions the temporal tagger should have access to resources in a modular way.

HeidelTime covers all types of temporal expressions, and additional rules can easily be added to HeidelTime's resources – a feature we used for extending the latest version of HeidelTime to better cover set expressions, for example. Due to HeidelTime's modular structure additional patterns and normalization resources can also be integrated easily, e.g., to extract additional implicit expressions.

5.2.3. Addressing Challenges 3 and 4

Challenges 1 and 2 do not require that a temporal tagger applies domain-dependent resources or strategies. However, challenge 3, i.e., the identification of the reference time for relative and underspecified temporal expressions, should be addressed depending on the domain of the documents that are to be processed. On the one hand, as described in Section 5.1, the reference time for underspecified and relative temporal expressions in news documents is often the document creation time (DCT). The same is true for such expressions in the colloquial and scientific corpora. On the other hand, narrative-style documents usually do not contain any relative or underspecified temporal expressions, for which the reference time is the DCT. In contrast, another temporal expression in the text has to be identified as reference time. Although identifying the reference time is sometimes difficult, a simple strategy is to use the previously mentioned expression of the required granularity as reference time. Thus, the reference time can be identified as depicted in Figure 3(a). Note, that there are some relative temporal expressions such as "two days later" for which the reference time has to be identified in the text independent of the domain of the document (context-dependent expressions). For these expressions, the previously mentioned ex-

pression can be used as reference time on all domains.

To identify the reference time of relative and underspecified expressions, HeidelTime realizes the strategy depicted in Figure 3(a). However, there are some temporal expressions in text documents, e.g., if they describe background information, that are not reliable candidates for reference times. To determine a temporal expression's eligibility to be a reference time may help in such cases. Currently, we are examining such strategies, e.g., to not use attributively occurring temporal expressions as reference times.

To normalize underspecified expressions, the next challenge is to identify the temporal relation to the reference time (challenge 4). If the reference time is the DCT, a promising approach is to identify the tense of the sentence. While past tense indicates that the relation to the reference time is "before", present tense and future tense indicate that the relation is "after". However, in some cases, there is no tense in the sentence and thus the relation has to be guessed. Then, we suggest a domain-dependent strategy. As described in the previous section, news are more likely to refer to past events, while SMS tend to refer to future events. If the reference time is not the DCT, which is the case if an underspecified expression occurs in narrative-style documents, a promising assumption is that the temporal expressions occur chronologically in the document. Note that this assumption is not made in general to all temporal expressions in a document but only concerns the underspecified expression that is under consideration and the previously mentioned expression, which is used as reference time. The strategies to determine the temporal relation to the reference time is depicted in Figure 3(b) and implemented in HeidelTime in the same way.

In summary, challenges 3 and 4 affect the normalization of temporal expressions and can be tackled using domain-dependent strategies, as it is done by HeidelTime.

5.2.4. Addressing Challenges 5 and 6

Challenge 5 (spelling variations, type errors, missing spaces) only occurs in colloquial text documents and thus only has to be tackled when processing documents from this domain. For spelling variations and word creations that refer to temporal expressions, e.g., "tmr" for "tomorrow", we suggest to add the synonyms to the pattern resources of the temporal tagger. A more difficult but also a frequently occurring challenge are type errors. We suggest to tackle this issue by searching for inexact patterns. Depending on the length of an expression that is to be matched, one could specify a threshold and calculate edit distances. If the edit distance is below the threshold, inexact matches could be extracted and normalized according to the edited expression. A third variation of challenge 5 are missing spaces between a temporal expression and the previous or next token. This could be tackled by removing the generally used constraint that a temporal expression has to begin and end with the beginning and ending of a token, respectively. To avoid too many false matches, one may want to validate that the whole token is not an existing word, e.g., by using a lexicon. This would avoid to wrongly match "May" in the expression "Mayonnaise", for instance.

Although we suggested strategies for all three types of non-standard language occurrences, HeidelTime only uses the strategy to identify spelling variations and word creations so far. This is realized in the following way: (i) We add a new language to HeidelTime (english-coll) by copying HeidelTime's English resources; (ii) the entries of all pattern resources are checked for synonyms using the noslang dictionary[9] that contains more than 5.000 entries for so-called Internet slang and acronym formulations that are often used in SMS as well; (iii) all synonyms are added to the pattern and normalization resources. In addition to setting the domain to "colloquial", one has to select "english-coll" as language when processing colloquial texts.

Finally, challenge 6 (unresolvable expressions) is a complex challenge occurring mainly in scientific data such as clinical trials. In order not to lose information by normalizing context-dependent relative expressions such as "two days later" to "XXXX-XX-XX" due to a missing reference time, we suggest to normalize such expressions according to a local time frame, i.e., a time point zero that may be defined in the document. To address this issue, we suggest to start with using the local semantics of temporal expressions as defined by Mazur and Dale (Mazur and Dale, 2011), e.g., that "one day later" is normalized to "+0000-00-01". However, beginning with the local semantics of the expression, we suggest to describe the semantics with respect to the local time frame of the document. Thus, in cases of chains of relative expressions the semantics can be accumulatively added. For example, a document about a clinical trial may contain the following text "...two days later ...one day later". Then "two days later" could be normalized to "TPZ+0000-00-02" and "one day later" to "TPZ+0000-00-03" referring to two and three days after time point zero (TPZ), respectively. In this way, we have annotated the expressions in the Time4SCI corpus (see Section 4.2). This strategy of normalizing context dependent relative expressions is realized by HeidelTime when selecting "scientific" as domain that is to be processed. Furthermore, similar to "english-coll", we add resources containing domain-dependent vocabulary and rules for the scientific domain ("english-sci").

6. Evaluating Domain-sensitive Strategies

When temporal tagging documents with HeidelTime, the user sets the language and the domain of the documents that are to be processed. In this section, we demonstrate the effectiveness of domain-sensitive temporal tagging based on this feature. We evaluate HeidelTime on the four described corpora applying the four domain-dependent settings (english/news, english/narrative, english-coll/colloquial, english-sci/scientific).

For the evaluation of the extraction and normalization tasks, the measures precision, recall, and f-score are widely used. In addition, there are five evaluation settings: *lenient* (extraction only, overlapping matches between gold standard and system annotations), *strict* (extraction only, exact matches), *value* (correct normalization of correctly extracted expressions), *len+val* (correct lenient extraction and correct value normalization), *str+val* (correct strict extraction and correct value normalization).

[9] http://www.noslang.com/dictionary/full/

corpus (domain)	strategy	lenient			strict			value			len+val			str+val		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
TimeBank (news)	**news**	**90.7**	91.5	**91.1**	**83.7**	**84.4**	**84.1**	**86.2**	**86.2**	**86.2**	**78.3**	**78.9**	**78.6**	**73.5**	**74.1**	**73.8**
	narrative	**90.7**	91.5	**91.1**	**83.7**	**84.4**	**84.1**	67.5	67.5	67.5	61.2	61.7	61.5	57.5	58.0	57.7
	colloquial	90.5	**91.7**	**91.1**	82.8	83.9	83.4	86.0	86.0	86.0	77.9	**78.9**	78.4	72.4	73.4	72.9
	scientific	**90.7**	91.5	**91.1**	83.0	83.7	83.4	81.2	81.2	81.2	73.7	74.3	74.0	69.0	69.6	69.3
WikiWars (narrative)	news	**93.9**	82.6	87.9	**86.0**	**75.7**	**80.5**	64.7	65.1	64.9	60.7	53.4	56.9	57.6	50.7	53.9
	narrative	**93.9**	82.6	87.9	**86.0**	**75.7**	**80.5**	**89.5**	**90.1**	**89.8**	**84.1**	**73.9**	**78.7**	**79.6**	**70.0**	**74.5**
	colloquial	93.3	**83.4**	**88.1**	84.3	75.3	79.6	64.1	64.5	64.3	59.8	53.5	56.5	56.0	50.0	52.8
	scientific	**93.9**	82.9	88.0	85.5	75.4	80.1	63.8	64.2	64.0	59.9	52.9	56.2	56.7	50.1	53.2
Time4SMS (colloquial)	news	99.3	85.2	91.7	**98.9**	84.8	91.3	**97.9**	**97.9**	**97.9**	**97.2**	83.4	89.8	**97.2**	83.3	89.7
	narrative	99.3	85.2	91.7	**98.9**	84.8	91.3	96.4	96.4	96.4	95.7	82.1	88.4	95.6	82.0	88.3
	colloquial	**99.4**	**91.1**	**95.1**	98.1	**90.0**	**93.9**	97.1	97.1	97.1	96.4	**88.5**	**92.3**	96.0	**88.1**	**91.9**
	scientific	99.3	85.3	91.8	98.8	84.8	91.3	97.8	97.8	97.8	**97.2**	83.4	89.8	97.1	83.3	89.7
Time4SCI (scientific)	news	**95.1**	55.0	69.7	76.2	44.1	55.8	74.4	74.4	74.4	70.8	40.9	51.9	67.6	39.1	49.5
	narrative	**95.1**	55.0	69.7	76.2	44.1	55.8	74.4	74.4	74.4	70.8	40.9	51.9	67.6	39.1	49.5
	colloquial	95.0	59.1	72.8	75.9	47.2	58.2	75.7	75.7	75.7	71.9	44.7	55.1	67.8	42.2	52.0
	scientific	**95.1**	**66.6**	**78.3**	**87.9**	**61.6**	**72.4**	**88.7**	**88.7**	**88.7**	**84.4**	**59.1**	**69.5**	**78.6**	**55.0**	**64.7**

Table 2: Evaluating HeidelTime using the different domain settings on the corpora of the four domains.

The results are presented in Table 2. On all corpora, the correct language/domain setting outperforms the other settings. Especially the quality of the normalization highly depends on the applied strategy. Note that the good results on Time4SMS with all strategies can be explained by the many occurring DCTs (1000), and that the colloquial setting still significantly outperforms the other settings. In summary, the results confirm our assumption that a domain-sensitive temporal tagger is necessary to achieve high quality results on different domains, especially for the normalization task.

7. Conclusions and Ongoing Work

In this paper, we analyzed the challenges of temporal tagging on documents of four domains, namely news-, narrative-, colloquial-, and scientific-style documents. For this, we developed gold standards for the colloquial and scientific domains and compared the annotated temporal expressions with those in existing corpora from the news and narrative domains. In addition, we suggested several strategies to address the identified challenges, described how they are realized in HeidelTime, and demonstrated their effectiveness performing a cross-domain evaluation.

Currently, we are implementing the suggested strategies to address typos and missing spaces. Furthermore, we analyze temporal expressions in literary texts. These often contain a local time frame, similar to the scientific documents. However, they may be very long and contain several unspecific time points that are relevant for normalizing other temporal expressions. We plan to adapt the scientific-style strategies to successfully process literary texts. In addition, we keep on improving HeidelTime and will publish evaluation results on the different corpora for the regularly updated publicly available versions of HeidelTime on our Website.

8. References

Tao Chen and Min-Yen Kan. 2011. Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus. Technical report, National University of Singapore.

Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. 2005. TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report, The MITRE Corporation.

Pawel Mazur and Robert Dale. 2007. The DANTE Temporal Expression Tagger. In *Language and Technology Conference '07*.

Pawel Mazur and Robert Dale. 2010. WikiWars: A New Corpus for Research on Temporal Expressions. In *EMNLP '10*, pages 913–922.

Pawel P. Mazur and Robert Dale. 2011. LTIMEX: Representing the Local Semantics of Temporal Expressions. In *FedCSIS'11*, pages 201–208.

Matteo Negri and Luca Marseglia. 2004. Recognition and Normalization of Time Expressions: ITC-irst at TERN 2004. Technical report, ITC-irst, Trento.

J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. 2003. The TIMEBANK Corpus. In *Corpus Linguistics '03*, pages 647–656.

James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Sauri. 2005. Temporal and Event Information in Natural Language Text. *Language Resources and Evaluation*, 39(2-3):123–164.

Estela Saquete and James Pustejovsky. 2011. Automatic Transformation from TIDES to TimeML Annotation. *Language Resources and Evaluation*, 45(4):495–523.

Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In *SemEval '10*, pages 321–324.

Jannik Strötgen and Michael Gertz. 2011. WikiWarsDE: A German Corpus of Narratives Annotated with Temporal Expressions. In *German Society for Computational Linguistics and Language Technology*, pages 129–134.

Jannik Strötgen and Michael Gertz. 2012. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, accepted for publication.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *SemEval '10*, pages 57–62.