Multilingual Normalization of Temporal Expressions with Masked Language Models

Lukas Lange¹ Jannik Strötgen¹ Heike Adel¹ Dietrich Klakow²

¹ Bosch Center for Artificial Intelligence, Renningen, Germany
² Spoken Language Systems (LSV), Saarland University, Saarbrücken, Germany
{Lukas.Lange,Jannik.Stroetgen,Heike.Adel}@de.bosch.com
dietrich.klakow@lsv.uni-saarland.de

Abstract

The detection and normalization of temporal expressions is an important task and preprocessing step for many applications. However, prior work on normalization is rule-based, which severely limits the applicability in real-world multilingual settings, due to the costly creation of new rules. We propose a novel neural method for normalizing temporal expressions based on masked language modeling. Our multilingual method outperforms prior rule-based systems in many languages, and in particular, for low-resource languages with performance improvements of up to 33 F₁ on average compared to the state of the art.

1 Introduction

Temporal tagging consists of the extraction of temporal expressions (TE) from texts and their normalization to a standard format (e.g., May '22: 2022-05). While there are deep-learning approaches for the extraction, temporal tagging as a whole is usually solved with highly specific rule-based systems, such as SUTime (Chang and Manning, 2012) or HeidelTime (Strötgen and Gertz, 2013). However, transferring rule-based methods to new languages or text domains requires a large manual effort to create rules specific to the target language. Although work on the automatic generation of rules for many languages (Strötgen and Gertz, 2015) exists, the rule quality typically does not match the high accuracy of hand-crafted rules.

In contrast to rule-based systems, neural networks are known for their ability to generalize to new targets, in particular, for cross- and multilingual applications (Rahimi et al., 2019; Artetxe and Schwenk, 2019). In the context of temporal tagging, recent works have shown promising results of neural networks for TE extraction in monolingual (Laparra et al., 2018) and multilingual settings (Lange et al., 2020) where a single neural model is trained on many languages at once. However, TE normalization remains challenging, and no solution for the normalization across languages exists yet.

We propose a new multilingual normalization method which can make use of labeled data from many languages by training a neural transformer model with a masked language modeling (MLM) objective. Thus, we adopt the MLM objective function for a new purpose: TE normalization.

To the best of our knowledge, this is the first work that uses neural networks for TE normalization. For this, as shown in Figure 1 and detailed below, we split the normalization task into two steps: normalization to a context-independent representation (CIR) using a slot-based masked language model (step 2) and anchoring given a reference time and further contextual cues (step 3).

The main contribution of this paper is our novel neural normalization method based on masked language modeling. For this, we create a large-scale multilingual dataset with weakly-supervised annotations of TEs and their normalized values in 87 languages. Our extensive set of experiments across 17 languages demonstrates that our multilingual method robustly works for many languages and outperforms the state of the art for multilingual temporal tagging, HeidelTime (Strötgen and Gertz, 2015), especially for low-resource languages by
more than 33 $F_1$ on average. Further, we explore different training and decoding strategies for our model. The code for our models and the weakly-supervised data is publicly available.\footnote{https://github.com/boschresearch/temporal-tagging-eacl}

2 Related Work

**TE Normalization.** Besides rule-based systems (Chang and Manning, 2012; Strötgen and Gertz, 2013), one normalization method for TEs are context-free grammars (Bethard, 2013; Lee et al., 2014) which are independent of the extraction method. However, they are even more language-specific than rule-based systems and hardly generalizable to new languages. Laparra et al. (2018) used a rule-based procedure for English TE normalization based on the SCATE format proposed by Bethard and Parker (2016). While their method could be extended to multilingual applications, no annotated data for other languages is available in the SCATE format, and it is mostly incompatible with the predominant TimeML (Pustejovsky et al., 2005) annotation format. Therefore, we will focus on the TimeML format in this work and present the first neural approach to TE normalization.

**Masked Language Modeling (MLM).** The MLM paradigm gains a lot of attention (Sun et al., 2021) due to popular language models like BERT (Devlin et al., 2019) This leads to active research on using MLM to solve further tasks like text classification (Brown et al., 2020), named entity recognition (Ma et al., 2021) and relation extraction (Han et al., 2021), also in low-resource languages (Hedderich et al., 2021). In this work, we adopt it to TE normalization for the first time.

3 Background on Temporal Tagging

Temporal tagging addresses the detection, classification and normalization of temporal expressions in unstructured texts — often following the TimeML specifications (Pustejovsky et al., 2005).

TimeML’s most important attributes are type (the class of an expression, e.g., DATE, TIME, DURATION or SET), and value (the normalized meaning of an expression, e.g., YYYY-MM-DD for specific days, such as 2022-05-01 for May 1, 2022). While some TEs contain all necessary information for the normalization, e.g., “May 1, 2022”, many expressions are incomplete w.r.t. the temporal information required for a normalization. An example is a relative expression like “yesterday” which needs an anchor point. Given the anchor point May 1, 2022, for example from the document creation time, “yesterday” should be annotated with type=DATE and value=2022-04-30.

Determining the anchor point can be challenging as it requires additional context information that could be given anywhere in the document. Therefore, systems for TE normalization, such as HeidelTime (Strötgen and Gertz, 2013), create an intermediate context-independent representation (CIR) of the value. In the syntax of HeidelTime, the expression yesterday would result in a CIR of UNDEF-last-day. Similarly, an underspecified expression, such as “May” would be represented with a CIR of UNDEF-year-05. Note that such a syntax for CIRs is language-independent. See Appendix A for more details. To determine the final value, the CIR needs to be anchored given, e.g., a reference date and further cues (such as tense information).

4 Approach

We propose to approach multilingual temporal tagging in three steps as shown in Figure 1: (1) Extraction of temporal expressions and their types using a multilingual sequence tagger; (2) Normalization of TEs to CIRs with our novel MLM-based normalization model; (3) Anchoring of CIRs given a reference time, e.g., using HeidelTime rules.

Our main contribution is a neural model for the second subtask, the normalization to a CIR. To the best of our knowledge, this has not been addressed with neural networks before. In this section, we detail all components of our approach. Information on the models that we apply for the first and third subtasks as well as an ablation study of directly predicting the normalized anchored value (without CIRs) are given in Section 5.

**Masked Language Modeling.** We model the task of assigning CIRs to temporal expressions as masked language modeling. In particular, we add TimeML annotations as inline information to the text sequences and mask the value field for prediction, e.g., “... <TIMEX3 type=’DATE’ value=’MASK’>yesterday</TIMEX3> ...”. Note that those annotations could be the ground-truth annotations when applying the model on gold temporal expressions or predicted temporal expressions when using the model in the 3-step pipeline as described above. In our experiments, we train a transformer model for CIR prediction using the
masked language modeling (MLM) objective.

**Slot-Based Value Representation.** Using only a single mask token for the whole CIR would require the model to store all possible CIRs in its vocabulary. Since it is not possible to enumerate, i.a., all possible dates, we model the CIRs as a fixed-length sequence of slots. In particular, we define 11 slots and use regular expressions to split the value field into slots in the training data. Figure 1 shows an example for the CIR “UNDEF-year-05” that is represented as the slots “ [PAD], [year], [05], [PAD], ..., [PAD]”. Details on the slots and regular expressions are given in Appendix A. To cover the full vocabulary of CIRs, we introduce 200 new tokens to the vocabulary of the language model.

**Curriculum Learning.** Our slot-based representation with 11 slots per CIR results in 11 masks. To train the model on this task, we apply curriculum learning in the first half of the training. In particular, we start with masking only a single slot of the CIR and steadily increase the number of masks up to the maximum of 11. For the second half of the training, masking is applied to all slots. We follow Devlin et al. (2019) and mask different parts of the input with different probabilities. In particular, we mask the value slots with a probability of 70%, annotated tokens with 15%, types with 10% and other text parts with 5%.

**Inference and Decoding.** For inference, we first add 11 masks (i.e., one per slot) to the input sentence. They serve as value placeholders that need to be predicted. Then, we use the masked language model to predict the most probable sequence of slots for the CIR. To decode the sequence, we apply sequential left-to-right decoding of all masks by iteratively decoding the left-most mask and replacing the mask with its predicted value until all masks are resolved. We compare this to two alternative decoding strategies: (i) decoding all masks simultaneously, (ii) training a conditional random field model that takes the logits as input and uses the Viterbi algorithm to determine the most probable sequence of predictions (Lafferty et al., 2001).

## 5 Experiments

This section describes our experiments and discusses the results. We compare our model to HeidelTime (Strötgen and Gertz, 2013), the current state of the art for multilingual temporal tagging.

For evaluation, we use the TempEval3 evaluation script (UzZaman et al., 2013) and report strict, relaxed and type $F_1$ for the extraction and value $F_1$ for our normalization experiments, respectively.

**Evaluation Data.** Our models are evaluated on gold-standard corpora in 17 languages. Details on the corpora are given in Appendix B.2. We divide the languages into high- and low-resource depending on whether manually created HeidelTime rules are available for the respective language.

**Training Data.** For training the normalization model, we create a large-scale weakly-supervised dataset covering 87 languages. Reasons are that (i) existing gold training data is too small to cover the wide range of different values and (ii) CIRs are not part of existing annotations. For all languages, we take the data from GlobalVoices (news-style documents) and Wikipedia (narrative-style documents), use Spacy for tokenization and HeidelTime for the annotation with temporal expressions. The number and quality of annotations is highly dependent on the amount of available data for that language and the quality of HeidelTime’s rules. Details on the weakly-supervised data are given in Appendix B.1.

**3-Step Pipeline for Temporal Tagging.** Both our temporal expression extraction and normalization models are based on the multilingual XLM-R transformer (Conneau et al., 2020).

We model the TE extraction as a sequence-labeling problem following Lange et al. (2020). For this, we convert the annotated corpora into the BIO format. For the monolingual setting (Mono), we train one model per language on the gold-standard resources if available or the weakly-supervised data otherwise. For the multilingual setting (Multi), we train a single model on the combined training resources of all languages.

For the normalization to CIRs, we train our proposed model with masked language modeling (see Section 4). In our experiments, we evaluate this model in combination with the multilingual extraction model (Multi+OUR) as well as in combination with the gold boundaries for temporal expressions (Gold+OUR) which serves as an upper bound.

---

2In our experiments, we compare our pre-defined slots to using subtokens from the language model tokenizer.

3The set of 87 languages is the intersection of languages covered by HeidelTime, our data and the XLM-R language model that we use for initializing our models.

4https://globalvoices.org/

5https://wikipedia.org/

6xlm-roberta-base with 270M parameters.
<table>
<thead>
<tr>
<th>Language</th>
<th>HeidelTime</th>
<th>Mono+OUR</th>
<th>Multi+OUR</th>
<th>Gold+OUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>de (N)</td>
<td>69.7</td>
<td>79.3</td>
<td>75.4</td>
<td>62.4</td>
</tr>
<tr>
<td>de (W)</td>
<td>88.5</td>
<td>94.3</td>
<td>89.0</td>
<td>84.8</td>
</tr>
<tr>
<td>en (N)</td>
<td>81.8</td>
<td>90.7</td>
<td>83.3</td>
<td>78.1</td>
</tr>
<tr>
<td>en (W)</td>
<td>90.6</td>
<td>94.3</td>
<td>90.6</td>
<td>94.3</td>
</tr>
<tr>
<td>es (N)</td>
<td>83.7</td>
<td>90.2</td>
<td>86.1</td>
<td>80.9</td>
</tr>
<tr>
<td>et (N)</td>
<td>42.4</td>
<td>57.4</td>
<td>51.3</td>
<td>44.0</td>
</tr>
<tr>
<td>fr (N)</td>
<td>85.6</td>
<td>90.6</td>
<td>82.3</td>
<td>73.3</td>
</tr>
<tr>
<td>hr (W)</td>
<td>93.3</td>
<td>95.8</td>
<td>94.6</td>
<td>85.7</td>
</tr>
<tr>
<td>it (N)</td>
<td>84.4</td>
<td>92.9</td>
<td>83.5</td>
<td>74.1</td>
</tr>
<tr>
<td>nl (N)</td>
<td>54.0</td>
<td>91.3</td>
<td>79.0</td>
<td>44.4</td>
</tr>
<tr>
<td>pt (N)</td>
<td>71.3</td>
<td>80.9</td>
<td>76.5</td>
<td>63.2</td>
</tr>
<tr>
<td>vi (W)</td>
<td>92.6</td>
<td>89.5</td>
<td>96.6</td>
<td>91.6</td>
</tr>
<tr>
<td>avg.</td>
<td>78.2</td>
<td>87.3</td>
<td>82.4</td>
<td>75.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>HeidelTime-auto</th>
<th>Mono+OUR</th>
<th>Multi+OUR</th>
<th>Gold+OUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ca (N)</td>
<td>28.1</td>
<td>62.8</td>
<td>61.1</td>
<td>43.6</td>
</tr>
<tr>
<td>el (W)</td>
<td>2.2</td>
<td>4.9</td>
<td>4.9</td>
<td>1.3</td>
</tr>
<tr>
<td>eu (N)</td>
<td>22.5</td>
<td>26.8</td>
<td>23.9</td>
<td>18.3</td>
</tr>
<tr>
<td>id (N)</td>
<td>19.7</td>
<td>54.7</td>
<td>44.5</td>
<td>40.1</td>
</tr>
<tr>
<td>pl (N)</td>
<td>18.8</td>
<td>27.2</td>
<td>16.5</td>
<td>11.2</td>
</tr>
<tr>
<td>ro (N)</td>
<td>3.2</td>
<td>19.5</td>
<td>16.7</td>
<td>5.5</td>
</tr>
<tr>
<td>ua (W)</td>
<td>1.6</td>
<td>2.6</td>
<td>2.2</td>
<td>1.2</td>
</tr>
<tr>
<td>avg.</td>
<td>12.7</td>
<td>28.4</td>
<td>24.3</td>
<td>17.3</td>
</tr>
</tbody>
</table>

Table 1: Detailed overview of our results for extraction (Str., Rel., Typ.) and normalization (Val.) per language for News and Wiki domains. The upper and lower parts display high- and low-resource languages, respectively.

For anchoring CIRs, we use rules similar to HeidelTime’s rules.\(^7\) In particular, anchor dates can be given by the document creation time or by previous temporal expressions (Strötgen and Gertz, 2016).

Results. Table 1 gives an overview of our experimental results. Multilingual extraction outperforms monolingual extraction, probably because the model is able to use knowledge from different languages. Our multilingual model achieves +2 \(F_1\) for high-resource and +51 \(F_1\) for low-resource languages compared to HeidelTime.

The normalization results are given in the Val. columns of Table 1. Our masked language model is matching HeidelTime’s performance rather close for high-resource languages and outperforms it for low-resource languages with an increase of 33 \(F_1\) points on average with our multilingual extraction model. Note that our models are multilingual, thus, we can use the same model for all languages.\(^8\) The upper bound of using gold extractions (Gold+OUR) shows that the extraction part still offers room for future improvements.

Note that HeidelTime with automatically created rules has a poor performance for some low-resource languages (el, ro, ua). This is similar to the observations by Grabar and Hamon (2019) who found that “[e]xploitation of this automatically built system produced no results when applied to the Ukrainian data.” For those languages, the automatic rule generation is not good enough in practice which emphasizes the need for multilingual systems like our model.

Ablation Studies. As our proposed model consists of multiple components, we now investigate their individual effects in more detail. The results for our ablation studies are given in Table 2.

First, we test different decoding strategies as described in Section 4. We find that sequential decoding works best. However, it also requires more computation time. A cheaper alternative with only minor performance decreases is the simultaneous decoding of all masks.

Second, we analyze the impact of different value representations by comparing our proposed approach with CIR and slot tokenization to (i) tokenization of values using the standard XLM-R
tokenizer instead of pre-defined slots (w/o OUR Slots), and (ii) training a model to directly predict the anchored value without CIRs in between (w/o OUR CIR). For (i), we find that our slot method has major advantages when processing narrative texts, such as Wikipedia, due to the higher amount of relative expressions (cf., Table 3 in Appendix B.3), that are tokenized into many subtokens (up to 34, instead of 11 when using our slots). For (ii), we add the document creation time to the input so that the model has more temporal information to predict the fully normalized value directly instead of a CIR. However, we find that current transformers are not able to correctly incorporate this information in a combined normalizing-anchoring step and mostly predict a memorized, incorrect value. Thus, using CIRs as an intermediate step is important for neural temporal tagging.

Third, we investigate the training strategy and training data. Our curriculum learning has advantages for low-resource languages as it reduces the training complexity which helps for the difficult adaptation to languages with few resources. Weakly-supervised training data is required, as the amount of gold-standard data is too small to train the MLM model. Finetuning the trained MLM model further on gold data (Weak+Gold) decreases performance slightly. Training the model on monolingual data only also decreases performance, highlighting the prospects of our multilingual approach.

Finally, we compare our models to an encoder-decoder model, i.e., an autoregressive language model that we adapt to TE normalization. For this, we follow the entity linking approach from De Cao et al. (2021) and train a BART encoder-decoder model (Lewis et al., 2020) for constrained decoding against a subspace of normalized TEs with our weakly-supervised data. We add the document creation time to the input, mark the extracted annotations and keep other TEs in the context, as in our other experiments. Given the gigantic amount of possible temporal expressions, e.g., there are roughly 32M seconds in a single year, we have to prune the search space to a reasonable size. Thus, we do not use time expressions of hour and smaller granularities and restrict the search space to years and months from 1 AD to 2100. Finer elements like weeks, days and daytimes are added for years between 2000 and 2026. We use durations for all defined units with numbers up to 10,000, e.g., 10,000 days. With this, we prune the search space to 1.4B terms which we store in a prefix tree. This results in an acceptable inference speed with BEAM search (5 beams). It takes roughly twice as long as our sequential MLM decoding. Note that this BART model has more parameters (400M) than the base version of XLM-R (270M) that we use in our model. The results are shown in the lower part of Table 2. We see, that our proposed MLM normalization model outperforms the BART model by a large margin. Nonetheless, the encoder-decoder model performs comparable to our model variant that directly predicts fully-normalized expressions. This clearly highlights the need for normalizing to CIRs before anchoring temporal expressions.

<table>
<thead>
<tr>
<th>Training Data</th>
<th>News de</th>
<th>en</th>
<th>Wiki de</th>
<th>en</th>
<th>Low-R. ca</th>
<th>eu</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o Curriculum</td>
<td>72.1</td>
<td>80.4</td>
<td>85.5</td>
<td>94.4</td>
<td>64.7</td>
<td>29.3</td>
</tr>
<tr>
<td>Encoder-Decoder Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monolingual</td>
<td>59.3</td>
<td>67.4</td>
<td>49.4</td>
<td>59.6</td>
<td>2.2</td>
<td>7.3</td>
</tr>
<tr>
<td>Multilingual</td>
<td>63.5</td>
<td>63.0</td>
<td>49.0</td>
<td>58.5</td>
<td>54.8</td>
<td>25.2</td>
</tr>
</tbody>
</table>

Table 2: Ablation study for our model components (Value $F_1$) on gold extractions. Low-R. stands for low-resource languages without gold training data.
Limitations

Our experiments are focused on Indo-European languages due to the lack of publicly available, labeled data points in other languages. Exceptions for which we could test zero-shot transfer were Basque, Estonian, Indonesian and Vietnamese. Even though, our model is working for these languages, it is not clear if the multilingual models transfer to all languages seen in the pre-training or by our weak supervision. The training of the multilingual models requires a considerable number of computational resources (up to 1.5 GPU days), which might not be available for all people/organizations. By publishing our model, we hope to lower the barrier for this kind of research by providing a pre-trained starting point. An in-depth error analysis to better understand which types of temporal expressions are well or less well covered in which language by our model was not performed. We are full of hope that such analyses can be tackled by users of our models who have the required language skills so that the analysis does not have to be limited to English.

Acknowledgments

We would like to thank the members of the BCAI NLP & NS-AI research group and the anonymous reviewers for their helpful comments.

References


Corina Forăscu and Dan Tufis. 2012. Romanian TimeBank: An annotated parallel corpus for temporal


Emmanouil I. Kapernaros. 2020. Extending the temporal tagger heideltim the for the greek language.


Jannik Strötgen, Ayser Armiti, Tran Van Canh, Julian Zell, and Michael Gertz. 2014. Time for more languages: Temporal tagging of arabic, italian, spanish,
and vietnamese. *ACM Transactions on Asian Language Information Processing*, 13(1).


A Slot Tokenization of CIRs

In this section, we describe our slot-based tokenization of the context-independent representation (CIR) of values as introduced in Section 3 and Section 4 of the main paper.

A.1 Overview of Slots

We use the following 11 slots to represent CIRs values.

- **SB:** This slot can contain BC information of years (e.g., as in BC4000 for the year 4000 BC) or the duration markers P and PT. Moreover, mathematical operations like PLUS are covered as used in relative expression involving offset computations (e.g., this-day-plus-2 for the day after tomorrow) and holiday names (EasterSunday).

- **SD1, SD2:** These slots are used to represent 4-digit year numbers (SD1 = 20 and SD2 = 22 for the year 2022) by splitting the 4-digit number into two 2-digit numbers. This helps to generalize to unseen years as fewer parameters have to be learned. In addition, we use SD1 to mark reference expressions like PAST_REF. For underspecified expressions like UNDEF-this-day, this is stored in SD1 and day in SD2. Moreover, SD1 and SD2 are used to store numbers of DURATION expressions.

- **SD3, SD4:** Analogously to SD1 and SD2 that are used to store year information, SD3 is used for months and SD4 for days.

- **ST1, ST2, ST3:** Temporal information from expressions of type TIME that are smaller than day granularity are stored in the ST slots. For example, the hour information of 24:00 and the daytime information, such as EV is stored in ST1. Information on minutes and seconds is stored in ST2 and ST3, respectively. Moreover, these slots are used to cover additional units in durations, such as in P1D2H (1 day and 2 hours).

- **SA1, SA2, SA3:** Finally, some CIRs include function calls which can be augmented with arguments that we store in the SA slots. For example, the argument 2 of this-day-PLUS-2 is stored in SA1. Other function calls are used to compute days with respect to holidays like EasterSunday or specific weekdays.

Note that slots can be optional depending on the temporal expression. For example, the value 2022 representing the year 2022 would only require SD1 and SD2. All other slots are set to a padding value [PAD] then which allows a fixed-sized representations of CIRs that can be predicted with our masked language model.

**Examples.** The following examples show temporal expressions, their corresponding CIRs and the tokenization into our slots. Note that there is no need to capture terms like UNDEF in our slots as the presence of words like this, next or last in a CIR implies the existence of UNDEF in the CIR.
This information can be reconstructed when obtaining a CIR from our slots. This also includes -- to separate numbers as in YYYY-MM-DD values, REF in reference expressions and T for time information.

We use the following format to give examples for our CIR conversion: Text → CIR → Slot Sequence

- Now ...
  → PRESENT_REF
  → SD1=PRESENT

- ... for 1000 days ...
  → P1000D
  → SB=P, SD1=10, SD2=00, SD4=D

- ... for one and a half day ...
  → P112H
  → SB=P, SD1=1, SD4=D, ST1=12, ST2=H

- ... in 1000 BC ...
  → BC1000
  → SB=BC, SD1=10, SD2=00

- ... on the morning of March 15, 2022 ...
  → 2022-03-15TMO
  → SD1=20, SD2=22, SD3=03, SD4=15, ST1=MO

- On March 15, ...
  → UNDEF-year=03-15
  → SD1=year, SD3=03, SD4=15

- ... the day after tomorrow ...
  → UNDEF-this-day-PLUS=2
  → SB=PLUS, SD1=THIS, SD2=day, SA1=2

- ... at Pentecost\(^{10}\) ...
  → UNDEF-year=00-00 funcDate...
    → Calc(EasterSunday(YEAR, 49))
  → SB=EasterSunday, SD1=year, SD2=00, SA1=49

A.2 Regular expressions

In the following, we will describe the six regular expressions used to split CIR values from HeidelTime outputs into our slots for the weakly-supervised training data.

Notation. For readability, we define the following groups to capture temporal units and other fixed names. Note that these are used across languages. For example, the German expression Montag would still be represented with monady.

\[^{10}\text{In Christian communities, the holiday of Pentecost is celebrated 49 days after Easter Sunday.}\]

\[\text{UNITS} = (H|D|DE|DT|M|C|Y|\]
\[\text{CE}|W|W|E|Qu|Q|S)\]
\[\text{UNITS}_F = (\text{day}|\text{month}|\text{year}\]
\[\text{decade|century|week|}
\[\text{weekend|quarter|}
\[\text{hour|minute|second}\]
\[\text{DAYTIME} = (NI|AF|M0|EV|MD|MI)\]
\[\text{SPECIAL} = (SP|SU|FA|AU|WI|}
\[H1|H2|Q1|Q2|Q3|Q4|H|Q)\]
\[\text{NAMES} = (\text{monday}|\text{tuesday}|\text{wednesday}|\text{thursday}|\text{friday}|\text{saturday}|\text{write}\]
\[\text{January|February|March|April|May|June|July|August|September|}
\[\text{October|November|December}\]

In the following, \(DX(n)\) marks the \(n\)-th group captured by the regular expression \(DX\).

D1: References. The first regular expression \(D1\) is used to capture simple reference expressions that refer to uncertain points in time.

\[\text{D1} = (\text{PRESENT}|\text{PAST}|\text{FUTURE})_\text{REF}\]

Slots: SD1=D1(1)

D2: Explicit Dates. The second regular expression \(D2\) detects explicit values that do not need further normalization, such as days in the YYYY-DD-MM format, e.g., 2022-03-15.

\[\text{D2} = \text{(BC)?(\d\d\d)?(XX)|((\d\d\d)|XX)?}\]
\[\text{(?::-(W)?(\d\d\d)?(XX)|SPECIAL))?}\]
\[\text{(?::-(\d\d\d)?(XX)|WE)?}\]
\[\text{(?::-((\d\d\d)?(X|DAYTIME)|XX)?}\]
\[\text{(?::((\d\d)))?}\]
\[\text{(?::(?:-(\d\d)?)})}\]

Slots: SB=D2(1), SD1=D2(2), SD2=D2(3), SD3=D2(5), SD4=D2(6), ST1=D2(7), ST2=D2(8), ST3=D2(9), D2(4)

P1: Durations. The third regular expression \(P1\) detects expressions of type DURATION, e.g., P1D2H. These are defined as \(P\text{<number><unit}\) for units of at least day granularity and \(PT\text{<number><unit}\) for smaller granularities. We capture up to two different units P1D2H (1 day and 2 hours) but ignore further units that are theoretically defined in the TimeML specifications but do not occur often in practice (in our datasets those did not occur at all).

\[\text{P1} = (\text{P|PT}(\d\d\d)?(\text{X|XX})}\]
\[\text{((\d\d\d)|X)?(\text{X|XX})+?}\text{(UNITS)?}\]
\[\text{(?::(\d\d)?(UNITS)?}\]

Slots: SB=P1(1), SD1=P1(2), SD2=P1(3), SD3=P1(4), SD4=P1(6), ST1=P1(5), ST2=P1(7)
D3: Relative Dates. While the previous regular expressions D1, D2 and P1 follow the TimeML specifications and capture fully normalized expressions, i.e., anchored values, the following regular expressions capture CIRs as used internally by HeidelTime. They represent relative expressions that need to be anchored.

D3 detects relative expressions with respect to a certain point in time, such as this-day-plus-2 (the day after tomorrow).

D3 = UNDEF-(this|next|last|REF|REFUNIT|REFDATE)?-?
(UNITS_F|SPECIAL)?-?
(NAMES|SPECIAL)|XX?(\d\d\d)?
(\?:-?\(\d\d\d\))?
(\?:-PLUS|MINUS|LESS)-(\d\d\d)?-?
\((?:?,\text{-?}(XX))?
\)?
\((?:T\(XX\))?)
\)?

Slots: SB=D3(5), SD1=D3(1),
SD2=D3(2), SD3=D3(3), SD4=D3(4),
ST1=D3(9), ST2=D3(10), ST3=D3(11),
SA1=D3(6), SA2=D3(7), SA3=D3(8)

D4: Relative Dates (coarse). D4 captures underspecified expressions like May that is missing year information and would be represented with the CIR UNDEF-year-05.

D4 = UNDEF-(year|decade|century)?
-?-(\d\d\d\d)|\?-(\d\d\d)?\(XX)?
\((?:,\text{-?}(\d\d\d))?)
\)?
\((?:T\(XX))?
\)?

Slots: SD1=D4(1), SD2=D4(2),
SD3=D4(3), SD4=D4(4),
ST1=D4(5), ST2=D4(6), ST3=D4(7)

D5: Holidays and functions. Finally, D5 covers special functions used by HeidelTime. These functions are used to compute days with respect to weekdays and moveable feasts like EasterSunday that refer to different days depending on the year. For example, the earliest possible date of Easter Sunday is March 22 and the latest is April 25 in the Gregorian calendar. The concrete date is then computed by an external function given a year.

D5 = (UNDEF-year|UNDEF-this-year|
UNDEF-arg-year|\(\d\d\d\d\)-\(\d\d\d\d\)-\(\d\d\d\d\)-
\(\d\d\d\d\)-\(\d\d\d\d\)-\(\d\d\d\d\d\d\)\-00 funcDateCalc\((\|
WeekdayRelativeTo|\|
EasterSundayOrthodox|\|
EasterSunday|)

\)

B Data Statistics

B.1 Weakly-Supervised Data

As detailed in Section 5, we create weakly-supervised data to train our normalization model, as the gold standard is too small and is not annotated with CIRs which are required by our method. For all languages, we take the data from GlobalVoices (news-style documents) and Wikipedia (narrative-style documents), use spacy for tokenization and our HeidelTime version that outputs CIRs for the annotation with temporal expressions. The sizes of our weakly-supervised data for each language are given in Table 4.

B.2 Gold-Standard Data

Detailed information on the datasets used in this paper (their languages, domains, sizes and references) are provided in Table 5. Note that all corpora come from the news domain except the WikiWars corpora that are based on Wikipedia articles.

B.3 Distribution of Explicit and Relative Values

The distribution of explicit and relative values has a large impact on the normalization performance of different models, as shown in our ablation study in Section 5. Exemplarily, we analyze their distribution in the German and English datasets for which we have data from two domains: News and Wikipedia. The results are given in Table 3. We see, that the Wikipedia corpora contain a much larger percentage of relative values as these articles often follow a narrative structure (cf., (Strögen and Gertz, 2016)).

C A Note on Adopting HeidelTime

In our experiments, we used a modified version of HeidelTime. First, we implemented a new UIMA collection reader based on spacy as an alternative to

13https://globalvoices.org/
the TreeTagger that has a restrictive license. This results in a slightly different sentence segmentation and tokenization, and, thus, minor differences in performance. For example, the original HeidelTime achieves 63.47 $F_1$ on the Portuguese test data, while our spacy version achieves 63.24 $F_1$ as one additional false positive expression was annotated due to different sentence boundaries. Second, we adapted HeidelTime to output its internal CIRs for the TimeML values, such that we can create our weakly-supervised training data.

The rather low performance of our models and HeidelTime for the high-resource languages Estonian (et) and Dutch (nl) can be explained by poor data quality. An inter-annotator agreement of 44 $F_1$ was reported for the Estonian corpus (Orasmaa, 2014), which is close to our results. The Dutch data was translated from English and automatically annotated via cross-lingual projections (Minard et al., 2016), which may reduce the annotation quality. Note, that only the first five sentences for each document were annotated in the Meantime corpora (it and nl). We restricted our evaluation to these annotated parts accordingly.

<table>
<thead>
<tr>
<th></th>
<th>De</th>
<th>En</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>67.1 / 32.9</td>
<td>52.3 / 47.7</td>
</tr>
<tr>
<td>Wiki</td>
<td>47.6 / 52.4</td>
<td>44.2 / 55.8</td>
</tr>
</tbody>
</table>

Table 3: Distribution of explicit / relative values according to HeidelTime by domains (in %).
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>de</td>
<td>870897</td>
<td>20</td>
<td>cy</td>
<td>14738</td>
<td>39</td>
<td>sq</td>
<td>5760</td>
</tr>
<tr>
<td>2</td>
<td>en</td>
<td>542087</td>
<td>21</td>
<td>da</td>
<td>12919</td>
<td>40</td>
<td>sr</td>
<td>4276</td>
</tr>
<tr>
<td>3</td>
<td>fr</td>
<td>284871</td>
<td>22</td>
<td>et</td>
<td>13948</td>
<td>41</td>
<td>ka</td>
<td>4247</td>
</tr>
<tr>
<td>4</td>
<td>ar</td>
<td>280446</td>
<td>23</td>
<td>sv</td>
<td>13705</td>
<td>42</td>
<td>el</td>
<td>4217</td>
</tr>
<tr>
<td>5</td>
<td>es</td>
<td>250871</td>
<td>24</td>
<td>ca</td>
<td>16747</td>
<td>43</td>
<td>he</td>
<td>4057</td>
</tr>
<tr>
<td>6</td>
<td>it</td>
<td>199236</td>
<td>25</td>
<td>bg</td>
<td>14550</td>
<td>44</td>
<td>sw</td>
<td>3979</td>
</tr>
<tr>
<td>7</td>
<td>nl</td>
<td>194944</td>
<td>26</td>
<td>fy</td>
<td>12852</td>
<td>45</td>
<td>ja</td>
<td>3696</td>
</tr>
<tr>
<td>8</td>
<td>ru</td>
<td>122884</td>
<td>27</td>
<td>pl</td>
<td>11283</td>
<td>46</td>
<td>br</td>
<td>3582</td>
</tr>
<tr>
<td>9</td>
<td>zh</td>
<td>105421</td>
<td>28</td>
<td>fa</td>
<td>11041</td>
<td>47</td>
<td>uz</td>
<td>3361</td>
</tr>
<tr>
<td>10</td>
<td>hr</td>
<td>50233</td>
<td>29</td>
<td>eu</td>
<td>10929</td>
<td>48</td>
<td>th</td>
<td>3162</td>
</tr>
<tr>
<td>11</td>
<td>ro</td>
<td>33545</td>
<td>30</td>
<td>ne</td>
<td>10750</td>
<td>49</td>
<td>cs</td>
<td>3096</td>
</tr>
<tr>
<td>12</td>
<td>vi</td>
<td>22048</td>
<td>31</td>
<td>ms</td>
<td>10017</td>
<td>50</td>
<td>ga</td>
<td>2799</td>
</tr>
<tr>
<td>13</td>
<td>af</td>
<td>21081</td>
<td>32</td>
<td>mg</td>
<td>9271</td>
<td>51</td>
<td>mn</td>
<td>2778</td>
</tr>
<tr>
<td>14</td>
<td>mk</td>
<td>19539</td>
<td>33</td>
<td>kk</td>
<td>8080</td>
<td>52</td>
<td>gd</td>
<td>2772</td>
</tr>
<tr>
<td>15</td>
<td>tr</td>
<td>19532</td>
<td>34</td>
<td>hi</td>
<td>7762</td>
<td>53</td>
<td>lt</td>
<td>2734</td>
</tr>
<tr>
<td>16</td>
<td>gl</td>
<td>17416</td>
<td>35</td>
<td>eo</td>
<td>7353</td>
<td>54</td>
<td>mr</td>
<td>2623</td>
</tr>
<tr>
<td>17</td>
<td>ca</td>
<td>16747</td>
<td>36</td>
<td>ur</td>
<td>6228</td>
<td>55</td>
<td>la</td>
<td>1876</td>
</tr>
<tr>
<td>18</td>
<td>bn</td>
<td>16284</td>
<td>37</td>
<td>hu</td>
<td>5871</td>
<td>56</td>
<td>uk</td>
<td>1673</td>
</tr>
<tr>
<td>19</td>
<td>cy</td>
<td>14738</td>
<td>38</td>
<td>sq</td>
<td>5760</td>
<td>57</td>
<td>hy</td>
<td>1642</td>
</tr>
<tr>
<td>20</td>
<td>bg</td>
<td>14550</td>
<td>39</td>
<td>sk</td>
<td>5172</td>
<td>58</td>
<td>ta</td>
<td>1556</td>
</tr>
<tr>
<td>21</td>
<td>et</td>
<td>13948</td>
<td>40</td>
<td>sr</td>
<td>4276</td>
<td>59</td>
<td>my</td>
<td>1103</td>
</tr>
<tr>
<td>22</td>
<td>sv</td>
<td>13705</td>
<td>41</td>
<td>ka</td>
<td>4247</td>
<td>60</td>
<td>ml</td>
<td>1079</td>
</tr>
<tr>
<td>23</td>
<td>id</td>
<td>13031</td>
<td>42</td>
<td>el</td>
<td>4217</td>
<td>61</td>
<td>kn</td>
<td>1029</td>
</tr>
<tr>
<td>24</td>
<td>da</td>
<td>12919</td>
<td>43</td>
<td>he</td>
<td>4057</td>
<td>62</td>
<td>fi</td>
<td>1017</td>
</tr>
<tr>
<td>25</td>
<td>fy</td>
<td>12852</td>
<td>44</td>
<td>sw</td>
<td>3979</td>
<td>63</td>
<td>oa</td>
<td>979</td>
</tr>
<tr>
<td>26</td>
<td>pl</td>
<td>11283</td>
<td>45</td>
<td>ja</td>
<td>3696</td>
<td>64</td>
<td>jv</td>
<td>968</td>
</tr>
<tr>
<td>27</td>
<td>fa</td>
<td>11041</td>
<td>46</td>
<td>br</td>
<td>3582</td>
<td>65</td>
<td>ky</td>
<td>926</td>
</tr>
<tr>
<td>28</td>
<td>eu</td>
<td>10992</td>
<td>47</td>
<td>uz</td>
<td>3361</td>
<td>66</td>
<td>is</td>
<td>804</td>
</tr>
<tr>
<td>29</td>
<td>eu</td>
<td>10992</td>
<td>48</td>
<td>th</td>
<td>3162</td>
<td>67</td>
<td>am</td>
<td>776</td>
</tr>
</tbody>
</table>

Table 4: Languages and the sizes of our weakly-supervised data.
<table>
<thead>
<tr>
<th>Corpus</th>
<th>Language</th>
<th>#Annotations (train / test)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Corpora only used for evaluation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRAUTS-DieZeit</td>
<td>German (de)</td>
<td>_ / 493</td>
<td>(Strötgen et al., 2018)</td>
</tr>
<tr>
<td>TempEval-3 (platinum)</td>
<td>English (en)</td>
<td>_ / 137</td>
<td>(UzZaman et al., 2013)</td>
</tr>
<tr>
<td>KOMPAS (test)</td>
<td>Indonesian (id)</td>
<td>_ / 192</td>
<td>(Mirza, 2016)</td>
</tr>
<tr>
<td>TimeBankCA</td>
<td>Catalan (ca)</td>
<td>_ / 1383</td>
<td>(Sauri, 2010)</td>
</tr>
<tr>
<td>EstTimeML</td>
<td>Estonian (et)</td>
<td>_ / 622</td>
<td>(Orasmaa, 2014)</td>
</tr>
<tr>
<td>EusTimeML</td>
<td>Basque (eu)</td>
<td>_ / 112</td>
<td>(Altuna et al., 2020)</td>
</tr>
<tr>
<td>Fr TimeBank</td>
<td>French (fr)</td>
<td>_ / 423</td>
<td>(Bittar et al., 2011)</td>
</tr>
<tr>
<td>Ro TimeBank</td>
<td>Romanian (ro)</td>
<td>_ / 151</td>
<td>(Forăscu and Tu fiyat, 2012)</td>
</tr>
<tr>
<td>PT-TimeBank (test)</td>
<td>Portuguese (pt)</td>
<td>_ / 151</td>
<td>(Costa and Branco, 2012)</td>
</tr>
<tr>
<td>WikiWars-EL (test)</td>
<td>Greek (el)</td>
<td>_ / 414</td>
<td>(Kapernaros, 2020)</td>
</tr>
<tr>
<td><strong>Corpora split into train and test sets</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meantime (IT)</td>
<td>Italian (it)</td>
<td>229 / 244</td>
<td>(Minard et al., 2016)</td>
</tr>
<tr>
<td>Meantime (NL)</td>
<td>Dutch (nl)</td>
<td>221 / 259</td>
<td>(Minard et al., 2016)</td>
</tr>
<tr>
<td>TempEval-3 (ES)</td>
<td>Spanish (es)</td>
<td>730 / 551</td>
<td>(UzZaman et al., 2013)</td>
</tr>
<tr>
<td>PolEval-2019</td>
<td>Polish (pl)</td>
<td>633 / 6011</td>
<td>(Kocon et al., 2019)</td>
</tr>
<tr>
<td>WikiWars</td>
<td>English (en)</td>
<td>1378 / 1251</td>
<td>(Mazur and Dale, 2010)</td>
</tr>
<tr>
<td>WikiWars-DE</td>
<td>German (de)</td>
<td>1510 / 684</td>
<td>(Strötgen and Gertz, 2011)</td>
</tr>
<tr>
<td>WikiWars-HR</td>
<td>Croatian (hr)</td>
<td>724 / 677</td>
<td>(Skukan et al., 2014)</td>
</tr>
<tr>
<td>WikiWars-UA</td>
<td>Ukrainian (ua)</td>
<td>454 / 2237</td>
<td>(Grabar and Hamon, 2019)</td>
</tr>
<tr>
<td>WikiWars-VI</td>
<td>Vietnamese (vi)</td>
<td>118 / 101</td>
<td>(Strötgen et al., 2014)</td>
</tr>
<tr>
<td><strong>Corpora only used for training</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRAUTS-Dolomiten</td>
<td>German (de)</td>
<td>388 / _</td>
<td>(Strötgen et al., 2018)</td>
</tr>
<tr>
<td>Meantime (EN)</td>
<td>English (en)</td>
<td>472 / _</td>
<td>(Minard et al., 2016)</td>
</tr>
<tr>
<td>TempEval-3 (train, en)</td>
<td>English (en)</td>
<td>1240 / _</td>
<td>(UzZaman et al., 2013)</td>
</tr>
<tr>
<td>PT-TimeBank (train)</td>
<td>Portuguese (pt)</td>
<td>1127 / _</td>
<td>(Costa and Branco, 2012)</td>
</tr>
<tr>
<td>WikiWars-EL (train)</td>
<td>Greek (el)</td>
<td>1496 / _</td>
<td>(Kapernaros, 2020)</td>
</tr>
</tbody>
</table>

Table 5: Overview of datasets and details on their usage as training (extraction-only) or evaluation data.