HEIDELTIME.HR: Extracting and Normalizing Temporal Expressions in Croatian

Luka Skukan, Goran Glavaš, Jan Šnajder

University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{luka.skukan, goran.glavas, jan.snajder}@fer.hr

Abstract

Temporal expression extraction and normalization are important for many NLP tasks and have been the topic of extensive research. While the majority of research on temporal expression extraction was performed for English, there has recently also been work on temporal processing for other languages. In this paper, we describe HEIDELTIME.HR, the Croatian resources for HeidelTime – a multilingual, cross-domain temporal expression tagger. HeidelTime recognizes temporal expressions in text and normalizes them according to the TIMEX3 annotation standard. We compile WikiWarsHr, a corpus of historical narratives in Croatian manually annotated for temporal expressions. On WikiWarsHr, HEIDELTIME.HR achieves results comparable to those originally achieved by HeidelTime on English texts, with F1-scores of 0.93 and 0.86 for expression extraction and normalization, respectively.

HEIDELTIME.HR: luščenje in normaliziranje časovnih izrazov v hrvaščini

Luščenje in normalizacija časovnih izrazov sta pomembna za raznovrstne naloge s področja računalniške obravnave naravnega jezika in sta bila predmet številnih raziskav. Medtem ko je bila večina raziskav luščenja časovnih izrazov opravljenih za angleščino, pa so bile v zadnjem času raziskave izvedene tudi za druge jezike. V prispevku opišemo HeidelTime.Hr, hrvaške vire za HeidelTime – večjezični in prekdomenski označevalec za časovne izraze. HeidelTime prepozna časovne izraze v besedilu in jih normalizira glede na standard za označevanje TIMEX3. Izdelamo WikiWarsHr, korpus zgodovinskih pripovedi v hrvaščini, ki je bil ročno označen za časovne izraze. Na WikiWarsHr doseže HeidelTime.Hr rezultate, primerljive s tistimi, ki jih je HeidelTime dosegal na angleških besedilih, z mero F 0,93 za luščenje in 0,86 za normalizacijo časovnih izrazov.

1. Introduction

The ability to extract and normalize temporal expressions in natural language texts is of major importance for natural language processing tasks, such as summarization and question answering, but also for reasoning about events and time in general. Temporal expression extraction is the task of identifying temporal expressions and their extent. The normalization task amounts to turning extracted temporal expressions into a fully specified value and formatting them according to some standard, including under-specified values.

While a number of temporal taggers are available, mostly for English and other major languages, a temporal expression tagger for Croatian does not yet exist. A new temporal expression tagger could be implemented, or an existing multilingual system could be adapted to work for Croatian. We chose the latter approach in this work, building on an existing and widely used framework.

In this paper, we describe HEIDELTIME.HR, the Croatian resources for the rule-based temporal expression tagger HeidelTime (Strötgen et al., 2013).[1] HeidelTime extracts and normalizes temporal expressions according to the TIMEX3 standard (Pustejovsky et al., 2003), and emerged as a winner in the TempEval-2 (Verhagen et al., 2010) and TempEval-3 (UzZaman et al., 2012) shared evaluation tasks. HeidelTime is a multilingual tagger, with resources been developed for English, German (Strötgen et al., 2013), Arabic, Italian, Spanish, Vietnamese (Strötgen et al., 2014a), French (Moriceau and Tannier, 2014), Chinese (Li et al., 2014), Dutch, and Russian. We have developed Croatian resources, which will be included in the next HeidelTime release.[2]

To develop and evaluate the tagger, we compiled WikiWarsHr, a corpus of historical narratives in Croatian manually annotated for temporal expressions. On this corpus, HEIDELTIME.HR achieves results comparable to those originally achieved by HeidelTime on English texts.

The structure of this paper is as follows. We describe the mechanisms of HeidelTime in Section 2. Section 3 describes the HEIDELTIME.HR resources. In Section 4, we describe the WikiWarsHr corpus and present the evaluation results. Section 5 concludes the paper.

2. HeidelTime tagger

The HeidelTime tagger extracts and normalizes temporal expressions according to the TIMEX3 standard (Pustejovsky et al., 2003). In TIMEX3, each temporal expressions is assigned a Type and a Value. A Type may be a *Date*, *Time*, *Duration* or *Set*. The Value corresponds to a temporal value, partially dependent on Type (e.g. a Date "*2014-10*" for October of 2014).

HeidelTime features a generic, language-independent core, written in Java, and a language-dependent part, the so-called language resources. A language resource consist of three sets: (1) expression resources , (2) normaliza-

[1] http://code.google.com/p/heideltime

[2] The HEIDELTIME.HR resources are also available from http://takelab.fer.hr/heideltimehr

> (DCT: June 21st 2014)
>
> The field of AI research was founded at a conference on the campus of Dartmouth College in the <TIMEX3 tid="t1" type="DATE" value="1956-SU">summer of 1956</TIMEX3>. <TIMEX3 tid="t2" type="DATE" value="2014">58 years later</TIMEX3>., we still haven't achieved many of the goals proposed there. Still, artificial intelligence has advanced and is <TIMEX3 tid="t3" type="DATE" value="2014-06-21">today</TIMEX3> a part of our daily lives without most of us knowing it.

Figure 1: Example of under-specification resolution.

tion resources, and (3) rule resources. Expression resources are regular expressions used for extraction temporal expressions from text, e.g., phrases for months, weekdays, numbers, etc. Normalization resources translate matched tokens to their canonical form, according to TIMEX3, by applying normalization mapping to extracted patterns (e.g., "*May*" → "*05*"). Finally, the rule resources combine the previous two resources to extract and normalize temporal expressions. These may be complemented with additional regular expressions to form more complex match-and-normalize rules, e.g., for discarding parts of extracted expressions or for adding a modifier ("*early*", "*middle*", etc.).

Normalization is performed both on fully specified expressions ("*June 28, 1995*") and relative temporal expressions ("*tomorrow*"). The latter are expressions that cannot be normalized without contextual information. Normalization of relative temporal expressions is performed by leaving the expressions under-specified and relying on HeidelTime's generic focus-tracking system to assign them a more specific value. For example, given a document creation time (DCT) of June 20th, 2014, the expression "*tomorrow*" might be resolved as "2014-06-21". This step is performed by taking into account the type of the document (narrative, news, scientific, or colloquial) and the tenses of the verbs used in the sentence containing the under-specified temporal expression. Either the DCT or a previously mentioned value can be used in under-specified expression normalization, depending on the document type and the normalization rule. An example of resolving under-specified dates using both DCT and current focus is shown in Fig. 1. Additionally, HeidelTime supports functionality extensions in form of text post-processors written as Java code. These allow for more verbose expression resolution, e.g., computing the date of lunar holidays such as Easter.

3. HEIDELTIME.HR

The task of developing resources for Croatian language consisted of developing three above-mentioned sets of resources. We next describe the resources and the development methodology.

3.1. Preprocessing

HeidelTime requires text to be pre-annotated with token, sentence and part-of-speech (POS) information. We used the CSTLemma lemmatiser (Jongejan and Haltrup, 2005) for token splitting and lemmatization,[3] and the HunPos part-of-speech tagger (Halácsy et al., 2007) to obtain the POS information. To integrate this functionality with HeidelTime, we wrote a Java wrapper that allows the tagger's engine to invoke it during pre-processing. HunPos and CSTLemma were previously trained to work with Croatian texts (Agić et al., 2013).

3.2. Resources

HEIDELTIME.HR resources are divided into several classes. The expression and normalization resources are divided into descriptive classes, according to their common roles in temporal constructs, with each normalization resource corresponding to an expression resource. Some examples include the *MonthLong* and *Timezone* resources. The rules are divided according to their semantics in the TIMEX3 standard into *Date*, *Time*, *Duration* and *Set* resources. Altogether, there are 199 rule resources for Croatian: 123 for dates, 37 for time, 24 for durations, and 15 for sets. This number is much larger than for English, but of comparable size to resources for other inflected languages, such as French, which has 157 rule resources (Moriceau and Tannier, 2014). Furthermore, as a highly inflected language, Croatian requires a large number of rule variations to account for the inflections. This issue could have been partially avoided by using lemmas instead of raw words. However, we chose not to do so for three reasons: (1) The implementation would be complex and time-consuming; (2) Due to the generic nature of the HeidelTime engine, lemmatization would have to be integrated system-wide, and the decision of whether to use lemmatization would have to be specified for each set of language-specific resources; (3) Errors in lemmatization would propagate into HeidelTime, decreasing its accuracy.

As an illustration, consider the following example of a complete HeidelTime extraction rule, which can be used to extract and normalize parts of seasons, such as "*ranog proljeća*" ("*early spring*"):

```
RULENAME="date_r9b",
EXTRACTION="%rePartWords_g1 %reSeason_g2",
NORM_VALUE="UNDEF-year-%normSeason(g2)",
NORM_MOD="%normPartWords(g1)"
```

The extraction part of the rule extracts expressions describing a specific part of something (e.g., "*early*", "*middle of*", etc.) and stores it as *group 1* (g_1), as well as an expression denoting a season (e.g., "*summer*"), which is stored as *group 2* (g_2). It leaves the year undefined as "UNDEF-year", which will be resolved by HeidelTime using the temporal context of the sentence. The "part word" in g_1 is normalized as part of the modifier (NORM_MOD), which makes the value more specific. The season, g_2, is combined with the determined year to get the temporal value of the expression. Assuming the inferred year is 2014, the given expression "*ranog proljeća*" would

[3] The lemmas produced by the CSTLemma lemmatiser are presently not used by the system, but may be integrated in the future (cf. Section 3.2.).

be normalized as *<TIMEX3 tid=”t1” value=”2014-SP” mod=”START”>ranog proljeća</TIMEX3>*.

3.3. Development methodology

We developed HEIDELTIME.HR in two phases. We first translated the existing English and German resources (Strötgen et al., 2013) into Croatian, wherever appropriate. We then used a data-driven approach to further develope and refine the resources, using a subset of manually-annotated Wikipedia corpus (cf. Section 4.1.) as a development set. The development set consists of ten Wikipedia articles of varying length, altogether containing 29,563 non-punctuation tokens and 677 temporal expressions. Usage examples for all TIMEX3 types of rule resources are given in Fig. 2.

(a) Službeno, američki angažman je završio u <TIMEX3 tid="t128" type="TIME" value="2010-08-31T17:00">utorak, 31. kolovoza, u 17:00 sati</TIMEX3>. Otprilike 50.000 vojnika je ostalo u Iraku do <TIMEX3 tid="t129" type="DATE" value="2011" mod="END"> kraja 2011.</TIMEX3>

(Officially, the American engagement ended on Tuesday, the 31st of October, at 5:00 PM. Around 50,000 soldiers stayed in Iraq until the end of 2011.)

(b) Rat nije bitnije promijenio granicu između dvije države, no cijena <TIMEX3 tid="t23" type="DURATION" value="P8Y">osmogodišnjeg</TIMEX3> ratovanja u ljudskim žrtvama i posljedicama po gospodarstvo je bila ogromna i za Irak i za Iran.

(The war did not result in major changes in the border between the two states, but the price of an eight-year war, in human lives and damage to the economy, was great for both Iraq and Iran.)

(c) Proizvodnja žita je opadala prosječno 3,5% <TIMEX3 tid="t55" type="SET" value="P1Y">godišnje</TIMEX3> između <TIMEX3 tid="t53" type="DATE" value="1978" >1978.</TIMEX3> i <TIMEX3 tid="t54" type="DATE" value="1990">1990.</TIMEX3> zbog borbi, nestabilnosti u seoskim područjima, duge suše i propale infrastrukture.

(The wheat production dropped, on average, 3.5% a year between 1978 and 1990, due to fighting, instability in rural areas, the long drought and the ruined infrastructure.)

Figure 2: Examples of Croatian documents tagged with HeidelTime.

When develeoping the rules, there were a few corner cases that we deliberately chose to ignore. More specifically, to not warrant a rule, an expression had to satisfy one of the following conditions:

1. A rule that would have been written to match the expression would be imprecise (i.e., result in more false positives than true positives). E.g., “*skoro* (“*soon/almost*”) is more often an adverb of degree than a temporal expression;

2. An expression is complex or unique, and therefore unlikely to appear in other documents, such as “*nedjelju oko 14,45 sati po srednjoeuropskome vremenu* (“*Sunday at about 2,45 PM according to Central European time*”);

3. An expression is very domain-specific and would potentially lead to a performance decrease across the board (e.g., references to the beginning or end of a particular war).

The rationale for the first condition is straightforward: recall would rise, but precision would plummet. The second and third conditions are meant to prevent overfitting. While including the specific rules would slightly increase the performance on the development set, it would not increase or could potentially decrease the performance on unseen data.

4. Evaluation

As part of this work, we have compiled WikiWarsHR, a corpus of Croatian Wikipedia manually annotated for temporal expressions. We used this corpus for the development and evaluation of HEIDELTIME.HR.

4.1. WikiWarsHR corpus

WikiWarsHR is inspired by the WikiWars corpus of Mazur and Dale (2010). While the content is similar (21 out of 22 articles detail the same wars as the original WikiWars corpus), the difference is that we chose to annotate WikiWarsHr using TIMEX3, a subset of the TimeML standard (Pustejovsky et al., 2003). The entire corpus contains almost 60,000 non-punctuation tokens and 1,440 temporal expressions in 22 articles. Two of these articles mostly contained historic (BC) temporal expressions, the processing of which is the newest addition to HeidelTime (Strötgen et al., 2014b) and which we have not yet implemented for Croatian. Therefore, we excluded these two articles, and divided the remaining 20 articles into the development and test set. This gave us a test set consisting of 10 articles, containing 21,644 tokens and 609 tagged temporal expressions. The articles are very diverse in length and temporal expression density, ranging from the minimum of 235 tokens and 12 tagged temporal expressions up to 9,722 tokens and 181 tagged temporal expressions. The WikiWarsHr corpus is freely available.[4]

4.2. Experimental setup

We computed the precision, recall, and F1-score for both expression extraction and normalization. The scores were computed on the expression level. We used two evaluation settings: *relaxed* and *strict*. In the relaxed setting,

[4] Available under the Creative Commons BY-NC-SA license from `http://takelab.fer.hr/wikiwarshr`

	Extraction			Type			Value		
Dataset & Tagger	P	R	F1	P	R	F1	P	R	F1
CroNER	0.83	0.54	0.66	0.82	0.54	0.65	–	–	–
HeidelTime (Development set)	0.95	0.97	0.96	0.94	0.96	0.95	0.86	0.88	0.87
HeidelTime (Test set)	0.94	0.96	0.95	0.93	0.95	0.94	0.86	0.88	0.87

Table 1: Tagger performance on WikiWarsHw corpus (relaxed match).

	Extraction			Type			Value		
Dataset & Tagger	P	R	F1	P	R	F1	P	R	F1
CroNER	0.26	0.17	0.21	0.26	0.17	0.21	–	–	–
HeidelTime (Development set)	0.93	0.95	0.94	0.93	0.95	0.94	0.86	0.88	0.87
HeidelTime (Test set)	0.92	0.93	0.93	0.91	0.93	0.92	0.85	0.87	0.86

Table 2: Tagger performance on WikiWarsHr corpus (strict match).

even partial matches are considered to be valid extraction and their normalization is scored. A partial match pertains to cases in which the tagged expression and the gold standard share at least one token. In the strict setting, only complete matches are considered correct. We computed the normalization scores for both the *Type* and *Value* properties of temporal expressions.

As a baseline, we evaluate CroNER (Glavaš et al., 2012) – a named entity recognition and classification system for Croatian – on the WikiWarsHr corpus. CroNER is capable of identifying temporal expressions belonging to *Date* and *Time* TIMEX3 types. As CroNER cannot normalize temporal expressions, we only evaluated expression extraction and type normalization. We measured the CroNER's performance on the entire WikiWars corpus (a union of the development and test set).

4.3. Results

Table 1 gives evaluation results for relaxed match. Extraction and normalization scores are high, particularly for the *Type*, with a negligible performance drop on the test set. Table 2 shows strict evaluation results for the two sets. The differences in the results compared to relaxed evaluation are almost negligible, with the drop in performance of 2% or less. This indicates that most errors are caused by errors in value normalization, rather than expression extraction.

Overall, the results are quite satisfying and comparable to those achieved by HeidelTime for English (Extraction 0.9; Type 0.82; Value 0.78) and Spanish (Extraction 0.9; Type 0.87; Value 0.85) (Strötgen et al., 2013).[5] However, part of this success can probably be attributed to the simpler nature of WikiWarsHr corpus in comparison to its English counterpart and a relatively large number of rules, many written primarily for the historical narratives.

4.4. Error analysis

As discussed above, most errors stem from value normalization. The few extraction errors are usually caused by unrecognized references to events or unique, large expressions. This is mostly due to the nature of the narratives – times relative to referenced events, implicitly switching focus between years, etc. Examples of both types of errors are given in Fig. 3.

Extraction error:

. . . nakon japanskog napada na Pearl Harbor . . .

(After the Japanese attack on Pearl Habor)

Normalization error:

Veljača, ožujak i travanj su bili relativno mirni mjeseci u usporedbi s krvavim studenim i siječnjom. . .

(The months February, March and April were relatively calm compared to the bloody November and January. . .)

Figure 3: Examples of tagging errors (expressions on which the errors occur are underlined).

Type	Errors	Occurrences	Error (%)
Date	85	1132	7.5
Time	1	23	4.3
Duration	1	50	2
Set	3	5	60

Table 3: Value normalization errors according to type.

Due to majority of errors originating from value normalization, we made a breakdown of normalization errors by expression type, on the union of the testing and development datasets. We considered only the expressions that have been correctly extracted (using strict evaluation) and had their type correctly normalized to match their counterparts in the gold standard. Table 3 shows that the largest number of errors stem from *Date* values, which also ac-

[5] Here we refer to the best results achieved on TempEval-3 datasets, obtained using tuned rulesets and relaxed matching.

count for the majority of temporal expressions in the corpus. Most of these errors are due to using a wrong focus point during normalization. Normalization of *Time* and *Duration* expression performs better, with a lower than 5% error rate. Value normalization performs poorly for *Set* values, with three out of five values normalized incorrectly. However, all three errors can be traced down to a systematic inconsistency: HeidelTime tagged all occurrences of "yearly" with the value "XXXX" (denoting "*every* year"), whereas the human annotators tagged it as "P1Y" (denoting "*once per* year"). In this case, however, the two tags are semantically equivalent.

5. Conclusion

We presented HEIDELTIME.HR, a resource we developed for temporal tagging of Croatian texts with the multilingual temporal tagger HeidelTime. We also described WikiWarsHr, a new Wikipedia-based corpus of Croatian historical narratives manually annotated for temporal expressions. On WikiWarsHr corpus, HEIDELTIME.HR achieves an F1-score of 0.93 and 0.86 for temporal expression extraction and normalization, respectively. This result is comparable to the result of HeidelTime for English.

Future extensions of the presented HiedelTime resources will include incorporating rules for historic dates, the newest addition to HeidelTime, into HEIDELTIME.HR. Furthermore, potential improvements in performance and ease of use could be achieved by adapting HeidelTime to work with lemmas instead of wordforms.

6. References

Ž. Agić, N. Ljubešić, and D. Merkler. 2013. Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*. Association for Computational Linguistics.

G. Glavaš, M. Karan, F. Šarić, J. Šnajder, J. Mijić, A. Šilić, and B. Dalbelo Bašić. 2012. CroNER: A State-of-the-Art Named Entity Recognition and Classification for Croatian. *Information Society*.

P. Halácsy, A. Kornai, and C. Oravecz. 2007. Hunpos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 209–212. Association for Computational Linguistics.

B. Jongejan and D. Haltrup. 2005. The CST Lemmatiser. *Center for Sprogteknologi, University of Copenhagen version*, 2.

H. Li, J. Strötgen, J. Zell, and M. Gertz. 2014. Chinese Temporal Tagging with HeidelTime. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 133–137.

P. Mazur and R. Dale. 2010. WikiWars: A new corpus for research on temporal expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 913–922. Association for Computational Linguistics.

V. Moriceau and X. Tannier. 2014. French resources for extraction and normalization of temporal expressions with HeidelTime. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

J. Pustejovsky, J. M. Castano, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, G. Katz, and D. R. Radev. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34.

J. Strötgen, J. Zell, and M. Gertz. 2013. HeidelTime: Tuning English and Developing Spanish Resources for TempEval-3. In *Proc. of Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 15–19.

J. Strötgen, A. Armiti, T. Van Canh, J. Zell, and M. Gertz. 2014a. Time for More Languages: Temporal Tagging of Arabic, Italian, Spanish, and Vietnamese. *ACM Transactions on Asian Language Information Processing (TALIP)*, 13(1):1.

J. Strötgen, T. Bögel, J. Zell, A. Armiti, T. V. Canh, and M. Gertz. 2014b. Extending heideltime for temporal expressions referring to historic dates. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2390–2397. European Language Resources Association (ELRA).

N. UzZaman, H. Llorens, J. Allen, L. Derczynski, M. Verhagen, and J. Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantcis (*SEM 2013)*. Association for Computational Linguistics.

M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62. Association for Computational Linguistics.