ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge

Michele Filannino, Gavin Brown, Goran Nenadic
The University of Manchester
School of Computer Science
Manchester, M13 9PL, UK
{m.filannino, g.brown, g.nenadic}@cs.man.ac.uk

Abstract

This paper describes a temporal expression identification and normalization system, ManTIME, developed for the TempEval-3 challenge. The identification phase combines the use of conditional random fields along with a post-processing identification pipeline, whereas the normalization phase is carried out using NorMA, an open-source rule-based temporal normalizer. We investigate the performance variation with respect to different feature types. Specifically, we show that the use of WordNet-based features in the identification task negatively affects the overall performance, and that there is no statistically significant difference in using gazetteers, shallow parsing and propositional noun phrases labels on top of the morphological features. On the test data, the best run achieved 0.95 (P), 0.85 (R) and 0.90 (F1) in the identification phase. Normalization accuracies are 0.84 (type attribute) and 0.77 (value attribute). Surprisingly, the use of the silver data (alone or in addition to the gold annotated ones) does not improve the performance.

1 Introduction

Temporal information extraction (Verhagen et al., 2007; Verhagen et al., 2010) is pivotal for many Natural Language Processing (NLP) applications such as question answering, text summarization and machine translation. Recently the topic aroused increasing interest also in the medical domain (Sun et al., 2013; Kovaćević et al., 2013).

Following the work of Ahn et al. (2005), the temporal expression extraction task is now conventionally divided into two main steps: identification and normalization. In the former step, the effort is concentrated on how to detect the right boundary of temporal expressions in the text. In the normalization step, the aim is to interpret and represent the temporal meaning of the expressions using TimeML (Pustejovsky et al., 2003) format. In the TempEval-3 challenge (UzZaman et al., 2012) the normalization task is focused only on two temporal attributes: *type* and *value*.

2 System architecture

ManTIME mainly consists of two components, one for the identification and one for the normalization.

2.1 Identification

We tackled the problem of identification as a sequencing labeling task leading to the choice of Linear Conditional Random Fields (CRF) (Lafferty et al., 2001). We trained the system using both human-annotated data (TimeBank and AQUAINT corpora) and silver data (TE3Silver corpus) provided by the organizers of the challenge in order to investigate the importance of the silver data.

Because the silver data are far more numerous (660K tokens vs. 95K), our main goal was to reinforce the human-annotated data, under the assumption that they are more informative with respect to the training phase. Similarly to the approach proposed by Adafre and de Rijke (2005), we developed a post-processing pipeline on top of the CRF sequence labeler to boost the results. Below we describe each component in detail.

2.1.1 Conditional Random Fields

The success of applying CRFs mainly depends on three factors: the labeling scheme (*BI*, *BIO*, *BIOE* or *BIOEU*), the topology of the factor graph and the quality of the features used. We used the *BIO* format in all the experiments performed during this research. The factor graph has been generated using the following topology: (w_0), (w_{-1}), (w_{-2}), (w_{+1}), (w_{+2}), $(w_{-2} \wedge w_{-1})$, $(w_{-1} \wedge w_0)$, $(w_0 \wedge w_{+1})$, $(w_{-1} \wedge w_0 \wedge w_{+1})$, $(w_0 \wedge w_{+1} \wedge w_{+2})$, $(w_{+1} \wedge w_{+2})$, $(w_{-2} \wedge w_{-1} \wedge w_0)$, $(w_{-1} \wedge w_{+1})$ and $(w_{-2} \wedge w_{+2})$.

The system tokenizes each document in the corpus and extracts 94 features. These belong to the following four disjoint categories:

- **Morphological**: This set includes a comprehensive list of features typical of Named Entity Recognition (NER) tasks, such as the word as it is, lemma, stem, pattern (e.g. *'Jan-2003': 'Xxx-dddd'*), collapsed pattern (e.g. *'Jan-2003': 'Xx-d'*), first 3 characters, last 3 characters, upper first character, presence of 's' as last character, word without letters, word without letters or numbers, and verb tense. For lemma and POS tags we use TreeTagger (Schmid, 1994). Boolean values are included, indicating if the word is lower-case, alphabetic, digit, alphanumeric, titled, capitalized, acronym (capitalized with dots), number, decimal number, number with dots or stop-word. Additionally, there are features specifically crafted to handle temporal expressions in the form of regular expression matching: cardinal and ordinal numbers, times, dates, temporal periods (e.g. *morning*, *noon*, *nightfall*), day of the week, seasons, past references (e.g. *ago*, *recent*, *before*), present references (e.g. *current*, *now*), future references (e.g. *tomorrow*, *later*, *ahead*), temporal signals (e.g. *since*, *during*), fuzzy quantifiers (e.g. *about*, *few*, *some*), modifiers, temporal adverbs (e.g. *daily*, *earlier*), adjectives, conjunctions and prepositions.

- **Syntactic**: Chunks and propositional noun phrases belong to this category. Both are extracted using the shallow parsing software MBSP[1].

- **Gazetteers**: These features are expressed using the BIO format because they can include expressions longer than one word. The integrated gazetteers are: male and female names, U.S. cities, nationalities, world festival names and ISO countries.

- **WordNet**: For each word we use the number of senses associated to the word, the first and the second sense name, the first 4 lemmas, the first 4 entailments for verbs, the first 4 antonyms, the first 4 hypernyms and the first 4 hyponyms. Each of them is defined as a separate feature.

The features mentioned above have been combined in 4 different models:

- **Model 1**: Morphological only
- **Model 2**: Morphological + syntactic
- **Model 3**: Morphological + gazetteers
- **Model 4**: Morphological + gazetteers + WordNet

All the experiments have been carried out using CRF++ 0.57[2] with parameters $C = 1$, $\eta = 0.0001$ and L2-regularization function.

2.1.2 Model selection

The model selection was performed over the entire training corpus. Silver data and human-annotated data were merged, shuffled at sentence-level (seed = 490) and split into two sets: 80% as cross-validation set and 20% as real-world test set. The cross-validation set was shuffled 5 times, and for each of these, the 10-fold cross validation technique was applied.

The analysis is statistically significant ($p = 0.0054$ with ANOVA test) and provides two important outcomes: (i) the set of WordNet features negatively affects the overall classification performance, as suggested by Rigo et al. (2011). We believe this is due to the sparseness of the labels: many tokens did not have any associated WordNet sense. (ii) There is no statistically significant difference among the first three models, despite the presence of apparently important information such as chunks, propositional

[1] http://www.clips.ua.ac.be/software/mbsp-for-python

[2] https://code.google.com/p/crfpp/

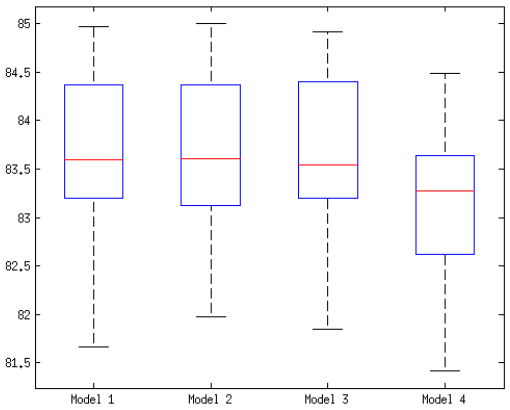

Figure 1: Differences among models using 5x10-fold cross-validation

noun phrases and gazetteers. The Figure 1 shows the box plots for each model.

In virtue of this analysis, we opted for the smallest feature set (Model 1) to prevent overfitting.

In order to get a reliable estimation of the performance of the selected model on the real world data, we trained it on the entire cross-validation set and tested it against the real-word test set. The results for all the models are shown in the following table:

System	Pre.	Rec.	$F_{\beta=1}$
Model 1	83.20	**85.22**	**84.50**
Model 2	**83.57**	85.12	84.33
Model 3	83.51	85.12	84.31
Model 4	83.15	84.44	83.79

Precision, Recall and $F_{\beta=1}$ score are computed using strict matching.

The models used for the challenge have been trained using the entire training set.

2.1.3 Post-processing identification pipeline

Although CRFs already provide reasonable performance, equally balanced in terms of precision and recall, we focused on boosting the baseline performance through a post-processing pipeline. For this purpose, we introduced 3 different modules.

Probabilistic correction module averages the probabilities from the trained CRFs model with the ones extracted from human-annotated data only. For each token, we extracted: (i) the conditional probability for each label to be assigned (*B*, *I* or *O*), and (ii) the prior probability of the labels in the human-annotated data only. The two probabilities are averaged for every label of each token. The list of tokens extracted in the human-annotated data was restricted to those that appeared within the span of temporal expressions at least twice. The application of this module in some cases has the effect of changing the most likely label leading to an improvement of recall, although its major advantage is making CRFs predictions less strict.

BIO fixer fixes wrong label sequences. For the BIO labeling scheme, the sequence *O-I* is necessarily wrong. We identified *B-I* as the appropriate substitution. This is the case in which the first token has been incorrectly annotated (e.g. *"Three/O days/I ago/I ./O"* is converted into *"Three/B days/I ago/I ./O"*). We also merged close expressions such as *B-B* or *I-B*, because different temporal expressions are generally divided at least by a symbol or a punctuation character (e.g. *"Wednesday/B morning/B"* is converted into *"Wednesday/B morning/I"*).

Threshold-based label switcher uses the probabilities extracted from the human-annotated data. When the most likely label (in the human-annotated data) has a prior probability greater than a certain threshold, the module changes the CRFs predicted label to the most likely one. This leads to force the probabilities learned from the human-annotated data.

Through repeated empirical experiments on a small sub-set of the training data, we found an optimal threshold value (0.87) and an optimal sequence of pipeline components (Probabilistic correction module, BIO fixer, Threshold-based label switcher, BIO fixer).

We analyzed the effectiveness of the post-processing identification pipeline using a 10-fold cross-validation over the 4 models. The difference between CRFs and CRFs + post-processing pipeline is statistically significant ($p = 3.51 \times 10^{-23}$ with paired T-test) and the expected average increment is 2.27% with respect to the strict $F_{\beta=1}$ scores.

2.2 Normalization

The normalization component is an updated version of NorMA (Filannino, 2012), an open-source rule-based system.

#	Training data	Identification						Normalization		Overall
		Strict matching			Lenient matching			Accuracy		
run	(post-processing)	Pre.	Rec.	$F_{\beta=1}$	Pre.	Rec.	$\tilde{F}_{\beta=1}$	Type	Value	score
1	Human&Silver (no)	78.57	63.77	70.40	97.32	78.99	87.20	88.99	77.06	67.20
2	Human&Silver (yes)	79.82	65.94	72.22	97.37	80.43	88.10	87.38	75.68	66.67
3	Human (no)	76.07	64.49	69.80	94.87	80.43	87.06	87.39	77.48	67.45
4	Human (yes)	78.86	**70.29**	**74.33**	95.12	**84.78**	**89.66**	86.31	76.92	**68.97**
5	Silver (no)	77.68	63.04	69.60	97.32	78.99	87.20	88.99	77.06	67.20
6	Silver (yes)	**81.98**	65.94	73.09	**98.20**	78.99	87.55	**90.83**	**77.98**	68.27

Table 1: Performance on the TempEval-3 test set.

3 Results and Discussion

We submitted six runs as combinations of different training sets and the use of the post-processing identification pipeline. The results are shown in Table 1 where the *overall score* is computed as multiplication between lenient $F_{\beta=1}$ score and the *value* accuracy.

In all the runs, recall is lower than precision. This is an indication of a moderate lexical difference between training data and test data. The relatively low *type* accuracy testifies the normalizer's inability to recognize new lexical patterns. Among the correctly typed temporal expressions, there is still about 10% of them for which an incorrect *value* is provided. The normalization task is proved to be challenging.

The training of the system by using human-annotated data only, in addition to the post-processing pipeline, provided the best results, although not the highest normalization accuracy. Surprisingly, the silver data do not improve the performance, both when used alone or in addition to human-annotated data (regardless of the post-processing pipeline usage).

The post-processing pipeline produces the highest precision when applied to the silver data only. In this case, the pipeline acts as a reinforcement of the human-annotated data. As expected, the post-processing pipeline boosts the performance of both precision and recall. We registered the best improvement with the human-annotated data.

Due to the small number of temporal expressions in the test set (138), further analysis is required to draw more general conclusions.

4 Conclusions

We described the overall architecture of ManTIME, a temporal expression extraction pipeline, in the context of TempEval-3 challenge.

This research shows, in the limits of its generality, the primary and exhaustive importance of morphological features to the detriment of syntactic features, as well as gazetteer and WordNet-related ones. In particular, while syntactic and gazetteer-related features do not affect the performance, WordNet-related features affect it negatively.

The research also proves the use of a post-processing identification pipeline to be promising for both precision and recall enhancement.

Finally, we found out that the silver data do not improve the performance, although we consider the test set too small for this result to be generalizable.

To aid replicability of this work, the system code, machine learning pre-trained models, statistical validation details and an online DEMO are available at: `http://www.cs.man.ac.uk/~filannim/projects/tempeval-3/`

Acknowledgments

We would like to thank the organizers of the TempEval-3 challenge. The first author would like also to acknowledge Marilena Di Bari, Joseph Mellor and Daniel Jamieson for their support and the UK Engineering and Physical Science Research Council for its support in the form of a doctoral training grant.

References

Sisay Fissaha Adafre and Maarten de Rijke. 2005. Feature engineering and post-processing for temporal expression recognition using conditional random fields. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, FeatureEng '05, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Ahn, Sisay Fissaha Adafre, and Maarten de Rijke. 2005. Towards task-based temporal extraction and recognition. In Graham Katz, James Pustejovsky, and Frank Schilder, editors, *Annotating, Extracting and Reasoning about Time and Events*, number 05151 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

Michele Filannino. 2012. Temporal expression normalisation in natural language texts. *CoRR*, abs/1206.2010.

Aleksandar Kovaćević, Azad Dehghan, Michele Filannino, John A Keane, and Goran Nenadic. 2013. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *Journal of American Medical Informatics*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. Timeml: Robust specification of event and temporal expressions in text. In *in Fifth International Workshop on Computational Semantics (IWCS-5*.

Stefan Rigo and Alberto Lavelli. 2011. Multisex - a multi-language timex sequential extractor. In *Temporal Representation and Reasoning (TIME), 2011 Eighteenth International Symposium on*, pages 163–170.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*.

Naushad UzZaman, Hector Llorens, James F. Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. *CoRR*, abs/1206.5333.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80, Prague.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 57–62, Stroudsburg, PA, USA. Association for Computational Linguistics.