HLT-FBK: a Complete Temporal Processing System for QA TempEval

Paramita Mirza
FBK, Trento, Italy
University of Trento
paramita@fbk.eu

Anne-Lyse Minard
FBK, Trento, Italy
minard@fbk.eu

Abstract

The HLT-FBK system is a suite of SVMs-based classification models for extracting time expressions, events and temporal relations, each with a set of features obtained with the NewsReader NLP pipeline. HLT-FBK's best system runs ranked 1st in all three domains, with a recall of 0.30 over all domains. Our attempts on increasing recall by considering all SRL predicates as events as well as utilizing event co-reference information in extracting temporal links result in significant improvements.

1 Introduction

QA TempEval is a continuation of the TempEval task series (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013), which shifts its evaluation methodology from temporal information extraction accuracy to temporal question-answering (QA) accuracy. However, the main task is the same as its predecessor tasks, which is to automatically annotate texts with temporal information following TimeML specification (Pustejovsky et al., 2003a).

This paper describes the HLT-FBK system submitted to QA TempEval. The system decomposes the task into three sub-tasks, i.e. temporal expression (timex) extraction, event extraction and temporal relation extraction. Each sub-task is formulated as a supervised classification problem using SVMs-based classifiers, which make use of the information acquired from the NewsReader[1] NLP pipeline.

[1] http://www.newsreader-project.eu

2 Data, Resources and Tools

The training data set is the TimeML annotated data released by the task organizers, which includes *TBAQ-cleaned* and *TE3-Platinum* corpora reused from the TempEval-3 task (UzZaman et al., 2013). We extended the training corpus for the timex extraction system with the TempEval-3 *silver* corpus.

The test data are 30 plain texts of *News*, *Wikipedia* and *Blogs* domains (10 documents each). For evaluating the system, 294 temporal-based questions and the test data annotated with entities relevant for the questions are used.

The resources used by the system to extract some features are **lists of temporal signals** extracted from the TimeBank corpus (Pustejovsky et al., 2003b) and a **list of nominalizations** extracted from the SPECIALIST Lexicon[2] distributed by the U.S. National Library of Medicine, which contains commonly occurring English words in addition to biomedical terms, with syntactic and morphological information. We extracted all nouns resulting from a nominalization. Other features come from the annotation of the **addDiscourse** tool (Pitler and Nenkova, 2009), which identifies discourse connectives and assigns them to one of the four semantic classes: *Temporal*, *Expansion*, *Contingency* and *Comparison*.

The **MorphoPro** module, part of the TextPro tool suite[3], is used to get the morphological analysis of each token in a text. The time expression nor-

[2] http://www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_001.html

[3] http://textpro.fbk.eu/

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 801–805,
Denver, Colorado, June 4-5, 2015.

malization sub-task is carried out by **TimeNorm**[4] (Bethard, 2013), a library for converting natural language expressions of dates and times into their normalized form.

The HLT-FBK system is a suite of classification models that have been built and applied using **YamCha**[5] (Kudo and Matsumoto, 2003), a text chunker using the Support Vector Machines (SVMs) algorithm. It supports the dynamic features that are decided dynamically during the classification, multi-class classification using either *one-vs-rest* or *one-vs-one* strategies, and *polynomial kernels*.

3 The End-to-end System

3.1 Pre-processing: NewsReader Pipeline

The data pre-processing was done using the NLP pipeline developed for the NewsReader project. The pipeline includes, amongst others, tokenization, part-of-speech tagging, constituency parser, dependency parser, named entity recognition, semantic role labeling (SRL) and event co-reference.[6]

3.2 Timex Extraction System

The task of recognizing the extent of a timex, as well as determining the timex type (i.e. DATE, TIME, DURATION and SET), is taken as a text chunking task. Since the timex extent can be a multi-token expression, we employ the IOB2 tagging to annotate the data, so each token will be classified into 9 classes: B-DATE, I-DATE, B-TIME, I-TIME, B-DURATION, I-DURATION, B-SET, I-SET and O (for other).

The classifier is built with *one-vs-one* strategy for multi-class classification. The features used to represent a token are token's text, lemma, part-of-speech (PoS) tag, chunk, named entity type (if any), and whether a token matches regular expression patterns for a time unit, part of a day, name of days, name of months, duration (e.g. *1h3'*), etc. In addition, all mentioned features for the preceding 4 and following 4 tokens, and the preceding 4 labels tagged by the classifier, are also included in the feature set.

For timex normalization, we decided to use TimeNorm. For English, it is shown to be the best performing system for most evaluation corpora (Llorens et al., 2012). We added pre- and post-processing rules in order to obtain the best normalized form.

3.3 Event Extraction System

Event detection is taken as a text chunking task, in which tokens have to be classified into two classes: EVENT (i.e. the token is included in an event extent) or O (for other). Then events are classified into one of the 7 TimeML classes (i.e. REPORTING, PERCEPTION, ASPECTUAL, I_ACTION, I_STATE, STATE and OCCURRENCE).

The classification models are built with *one-vs-rest* strategy for multi-class classification. For both event extent identification and event classification tasks we use various features to represent each token. The classic features are token's lemma, PoS tag, and entity type (if the token is part of a named entity or a time expression). Other features that are more specific for the task include: verb's tense and polarity[7], whether the token is annotated as predicate by the SRL module, whether it is part of an event co-reference chain and whether it is in the nominalization list. In addition, all mentioned features for the preceding 4 and following 4 tokens, and the preceding 4 labels tagged by the classifier, are also considered as features.

Specifically for event classification, additional features are used: token's chunk, whether the token is part of a temporal discourse connective, whether a verb is the main verb of the sentence (*root* verb), the predicate for which the token is part of a participant and its semantic role (e.g. Arg0, Arg1), and finally whether the token is in an event extent (annotated in the previous step).

We submitted two different runs:

- **Run 1** (*ev1*) Two classifiers are used as described above.
- **Run 2** (*ev2*) We consider all predicates identified by the SRL module as events. We then used a classifier to determine the class of each event.

[4] http://github.com/bethard/timenorm

[5] http://chasen.org/~taku/software/yamcha/

[6] More information about the NewsReader pipeline, as well as a demo, are available on the project website http://www.newsreader-project.eu/results/.

[7] The tense, aspect and polarity attributes of events, as defined in TimeML, are obtained through manually written rules based on the morphological analysis produced by MorphoPro.

3.4 Temporal Relation Extraction System

The temporal relation extraction system extracts temporal relations (TLINKs) holding between two events or between an event and a time expression. We consider all combinations of event/event and event/timex pairs within the same sentence (in a forward manner[8]), and pairs of main events (*root* verbs) of consecutive sentences, as candidate temporal links.

Given an ordered pair of entities (e_1, e_2), either event/event or event/timex pair, the classifier has to assign a label, i.e one of the 13 TimeML temporal relation types. However, we simplified the considered temporal relation types to better fit the QA TempEval task description and to deal with the unbalanced training data as follows: *(i)* IDENTITY and DURING are mapped to SIMULTANEOUS; *(ii)* IBEFORE/IAFTER are mapped to BEFORE/AFTER;[9] and *(iii)* INCLUDES, BEGINS and ENDS are converted to their inverse counterparts (IS_INCLUDED, BEGUN_BY and ENDED_BY, resp.) by exchanging the order of entities in the pair. In the end, we only consider 6 temporal relation types (i.e. SIMULTANEOUS, BEFORE, AFTER, IS_INCLUDED, BEGUN_BY and ENDED_BY).

The classification models for event/event and event/timex pairs are built with *one-vs-one* strategy for multi-class classification. The overall approach is largely inspired by an existing work for classifing temporal relations (Mirza and Tonelli, 2014). The implemented features are as follows:

String and grammatical features. Tokens, lemmas, PoS tags and chunks of e_1 and e_2, along with a binary feature indicating whether e_1 and e_2 in an event/event pair have the same PoS tags.

Textual context. Sentence distance (e.g. 0 if e_1 and e_2 are in the same sentence) and entity distance inside a sentence (i.e. the number of entities occurring between e_1 and e_2).

Entity attributes. Event attributes (*class*, *tense*, *aspect* and *polarity*) taken from the output of the event extraction module, and the timex attribute (*type*) obtained from the timex extraction module of e_1 and e_2; a binary feature to represent whether the timex in an event/timex pair is the document creation time; and four binary features to represent whether e_1 and e_2 in an event/event pair have the same event attributes or not. We also include as features the PoS chain of VP chunks containing events (e.g. VHZ-VBN-VVG for *has been [raining]*$_{e_1}$, VM-VVB for *would [send]*$_{e_2}$), which captures tense and aspect, as well as modality information of the event.

Dependency information. Dependency path existing between e_1 and e_2, and binary features indicating whether e_1/e_2 is the *root* verb.

Temporal signals. Tokens of temporal signals occurring around e_1 and e_2 and their positions with respect to e_1 and e_2 (i.e. *before/after* e_1, *before/after* e_2, or at the beginning of the sentence).

Temporal discourse connectives. We take into account discourse connectives belonging to the *Temporal* class, acquired from the *addDiscourse* tool. Similar to temporal signals, tokens of connectives occurring in the textual context of e_1 and e_2, and their position with respect to e_1 and e_2, are used as features. These features are only relevant for event/event pairs.

There are two variations of system submitted:

- **Run 1** (*trel1*) We incorporate pre-processing rules based on timex pattern matching (e.g. *from...to...*, *between...and...*), to recognize event/timex pairs of BEGUN_BY and ENDED_BY types, which are not well represented in the training corpus.
- **Run 2** (*trel2*) Similar as Run 1, however, we also incorporate the event co-reference information obtained from the NewsReader pipeline. Whenever two events co-refer, the event/event pair is excluded from the classifier, and automatically labelled SIMULTANEOUS.

4 Results

We submitted 4 system runs, i.e. the combinations of 2 system runs for event extraction (*ev1* and *ev2*) and 2 system runs for temporal relation extraction (*trel1* and *trel2*). Table 1 shows HLT-FBK system results in terms of coverage, precision, recall and F1-score for the three considered domains; recall is the main evaluation metric used to rank the systems.

[8] For example, for a sentence "...ev_1...tmx_1...ev_2...", the candidate pairs are (ev_1, tmx_1), (ev_1, ev_3) and (ev_2, tmx_1).

[9] Because event pairs of IBEFORE/IAFTER types are too scarce as training examples, and they are by definition specific types of BEFORE/AFTER.

	News				Wikipedia				Blogs				All domains
	Cov	**P**	**R**	**F1**	**Cov**	**P**	**R**	**F1**	**Cov**	**P**	**R**	**F1**	**R**
ev1-trel1	0.29	0.59	0.17	0.27	0.29	0.55	0.16	0.25	0.32	0.57	0.18	0.28	0.17
ev1-trel2	0.55	0.43	0.23	0.30	0.50	0.52	0.26	0.35	0.43	0.43	0.18	0.26	0.23
ev2-trel1	0.36	0.56	0.20	0.30	0.29	0.58	0.17	0.26	0.29	0.47	0.14	0.21	0.17
ev2-trel2	0.69	0.43	**0.29**	0.35	0.58	0.62	**0.36**	0.46	0.58	0.34	**0.20**	0.25	**0.30**

Table 1: HLT-FBK system results in terms of coverage (Cov), precision (P), recall (R) and F1-score (F1).

	News					Wikipedia					Blogs				
		Answered		Unknown			Answered		Unknown			Answered		Unknown	
	Q	**Cor**	**Inc**	**Ent**	**Rel**	**Q**	**Cor**	**Inc**	**Ent**	**Rel**	**Q**	**Cor**	**Inc**	**Ent**	**Rel**
ev2-trel1	99	20	16	17	46	130	22	16	48	44	65	9	10	22	24
ev2-trel2	99	29	39	16	15	130	47	29	48	6	65	13	25	22	5

Table 2: HLT-FBK system results in terms of number of answered questions, correctly (Cor) and incorrectly (Inc), and unanswered questions because of unknown entities (Ent) and unknown relations (Rel).

	News		Wikipedia		Blogs	
	ev	tx	ev	tx	ev	tx
ev1	0.72	0.83	0.81	0.59	0.68	0.35
ev2	0.80	0.83	0.84	0.54	0.70	0.35

Table 3: HLT-FBK system results in terms of recall on identifying events (ev) and timexes (tx) with strict match.

The best results are achieved with the combination of *ev2* and *trel2*, which significantly outperformed other participating systems and reported off-the-shelf systems (not optimized for the task), i.e. *CAEVO* with 0.17 and 0.18 recall scores on News and Blogs respectively, and *TIPSem* with 0.19 recall on Wikipedia.

Table 2 compares *trel1* and *trel2* runs, in terms of the number of answered questions (correctly and incorrectly) and unanswered questions (due to unknown entities and non-established/unknown relations). Meanwhile, Table 3 compares *ev1* and *ev2* in terms of recall scores on identifying EVENT and TIMEX3 tags, with the annotated test data as the gold standard.[10] Both results give more insight on the question answering-based evaluation.

5 Discussion

The timex extraction system performs well on News texts, but not on texts from Wikipedia and Blogs (see Table 3). Our error analysis shows that many time expressions in Wikipedia texts are not represented in the training corpus (e.g. *4th millennium BCE*).

Considering all SRL predicates as events (*ev2*) improves the recall on identifying relevant events (see Table 3), but lowers the precision on answering the questions (except for Wikipedia, in which the precision is also improved, see Table 1). In this task, the focus is on the recall and as expected the best results are obtained by the system with the best recall (*ev2*).

For temporal relation extraction, using event coreference information (*trel2*) reduces the number of unknown relations (Rel) down by 77% in average for all domains (see Table 2). Hence, the recall scores increase significantly as shown in Table 1, especially for the Wikipedia domain with almost 20% improvement.

Our attempts on improving the overall performance by increasing the recall (*ev2* and *trel2* runs) work well on News and Wikipedia, shown by improving F1-scores. This unfortunately does not hold for Blogs, since the precision is greatly compromised while the recall is only slightly improved.

In general, the system performs best on News and Wikipedia texts, but not so well on informal Blogs texts. This difference can be due to the fact that our systems, as well as most of the pipeline's modules, are trained using the corpus of formal news texts. Moreover, Blogs texts contain orthographic errors, a lot of punctuation signs, etc. and their preprocessing with the pipeline do not run well.

[10]The gold standard only contains the annotated entities relevant for answering the set of questions. For this reason, we computed only the recall.

Acknowledgments

The research leading to this paper was partially supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404).

References

Steven Bethard. 2013. A Synchronous Context Free Grammar for Time Normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 821–826, Seattle, Washington, USA.

Taku Kudo and Yuji Matsumoto. 2003. Fast Methods for Kernel-based Text Analysis. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 24–31, Stroudsburg, PA, USA.

Hector Llorens, Leon Derczynski, Robert Gaizauskas, and Estela Saquete. 2012. TIMEN: An Open Temporal Expression Normalisation Resource. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3044–3051, Istanbul, Turkey.

Paramita Mirza and Sara Tonelli. 2014. Classifying Temporal Relations with Simple Features. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 308–317, Gothenburg, Sweden.

Emily Pitler and Ani Nenkova. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 13–16, Stroudsburg, PA, USA.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003b. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656, Lancaster, March.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 1–9, Atlanta, Georgia, USA.

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 Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62.