School Research Student Symposium - Manchester, 30th October 2012

Temporal Information Extraction in the Clinical Domain

Michele Filannino
filannim@cs.man.ac.uk

Supervisors: Goran Nenadic, Gavin Brown
School of Computer Science, University of Manchester, UK

http://www.cs.man.ac.uk/~filannim

Motivation

Temporal aspects are important to **organise information**. They allow people to filter information and infer temporal **flows of events**. **Clinical documents**, in particular, contain patient's clinical events often not chronologically presented in text.

A temporal expression refers to a natural language phrase that denotes a temporal entity such as an interval or a time instant. They are annotated using **ISO-TimeML** [1] standard.

Despite its growing interest, there are **no publicly available** temporal information extraction systems for clinical data yet and the generic ones perform poorly because of the specificity of the clinical sub-language.

Problem

Identification:

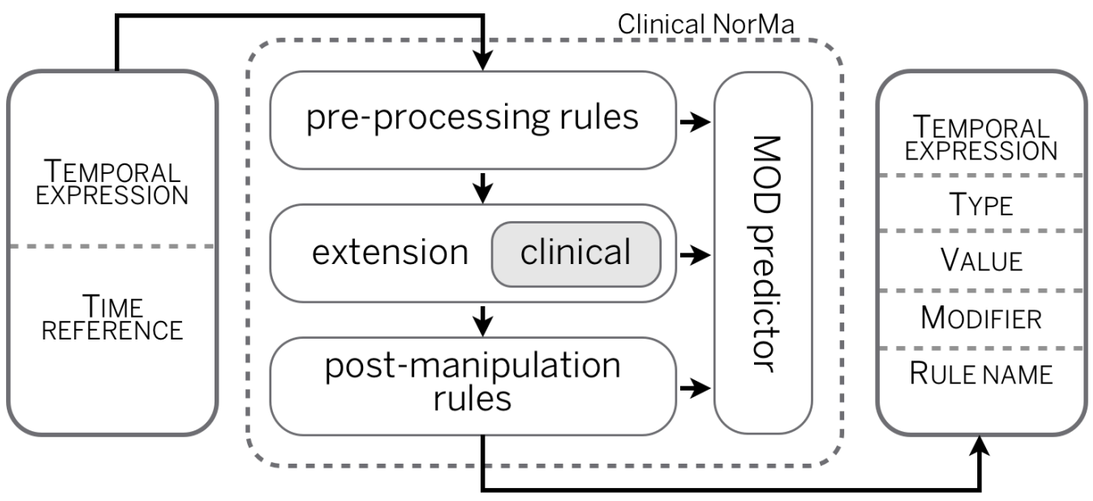
She delivered a 3680 gram male infant on 10/12/2004 at 10:17 pm with apgar scores of 9 and 9 at approximately one and five minutes respectively at 40.0 weeks gestation via spontaneous vertex vaginal delivery. She was discharged three days after in the morning.

Normalisation:

```
<TIMEX3 tid="t1" value="2004-10-12T22:17" type="TIME" />
<TIMEX3 tid="t2" value="PT1M" type="DURATION" mod="APPROX"/>
<TIMEX3 tid="t3" value="PT5M" type="DURATION" mod="APPROX"/>
<TIMEX3 tid="t4" value="P40W" type="DURATION" />
<TIMEX3 tid="t5" value="2004-10-15TMO" type="TIME" mod="START"/>
```

Architecture

Clinical NorMa is a hybrid system involving **dictionary-driven regular expression rules and** a **machine learning**-based component for the modifier predictor.



Rule example

```
pattern = re.findall('^(?:the |her  |his |their )?([0-9][0-9]*)(?:st|nd|rd
|th) (?:post-|post|day)? ?(?:pod| operative |op| hospital
|hsp|day|hd)(?:ly)? (?:day|night|afternoon)?$', raw_expression)
if pattern:
        value = add_date(reference_date , int(pattern[0]) )
        return expression, 'DATE' , value, 'postoperative_literals3'
```

temporal expression | type | ISO-8601 representation (value) | rule name

Results

System	Type	Value	Modifier
Edinburgh	0.84	0.63	-
HeidelTime	0.96	0.85	-
KUL	0.91	0.55	-
TERSEO	**0.98**	0.65	-
TipSem	0.92	0.65	-
TRIOS	0.94	0.76	-
NorMA	0.95	**0.86**	-

TempEval-2 test set: General domain (20 docs)

Systems	Type	Value	Modifier
Maximum score	**0.89**	**0.73**	**0.89**
Clinical NorMA [2]	0.85	0.70	0.83
Medium	0.78	0.60	0.79

i2b2 2012 test set: Clinical domain (120 docs)

Future Work

- Investigate the identification phase, combining **CRFs** with **semi-supervised** approach.
- Formalise the **normalisation** phase as a machine learning problem, tailoring a methodology to adjust general domain normaliser to a specific domain and testing it in the clinical

Acknowledgements

I am grateful to **my supervisors** for their precious advices and suggestions. I am also grateful to **Dr. Naushad UzZaman** from the University of Rochester for his support and appreciation about my research. I would also like to acknowledge the support of UK **Engineering and Physical Science Research Council** in the form of doctoral training grant.

[1] J. Pustejovsky, K. Lee, H. Bunt, and L. Romary. ISO-TimeML: An international standard for semantic annotation. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[2] A. Kovacevic, A. Dehghan, M. Filannino, J. Keane, and G. Nenadic, Extraction of events, temporal expressions and relations from clinical narratives using rules and machine-learning . Journal of the American Medical Informatics Association (2012)