NTCIR-11
with the 6th International Workshop on Evaluating Information Access,
Tokyo, 9-12 December 2014

Using machine learning to predict temporal orientation of search engines' queries in the Temporalia challenge

Michele Filannino[1]* and Goran Nenadic[1]

[1] School of Computer Science, University of Manchester, United Kingdom

* Presenting author, e-mail: filannim@cs.man.ac.uk

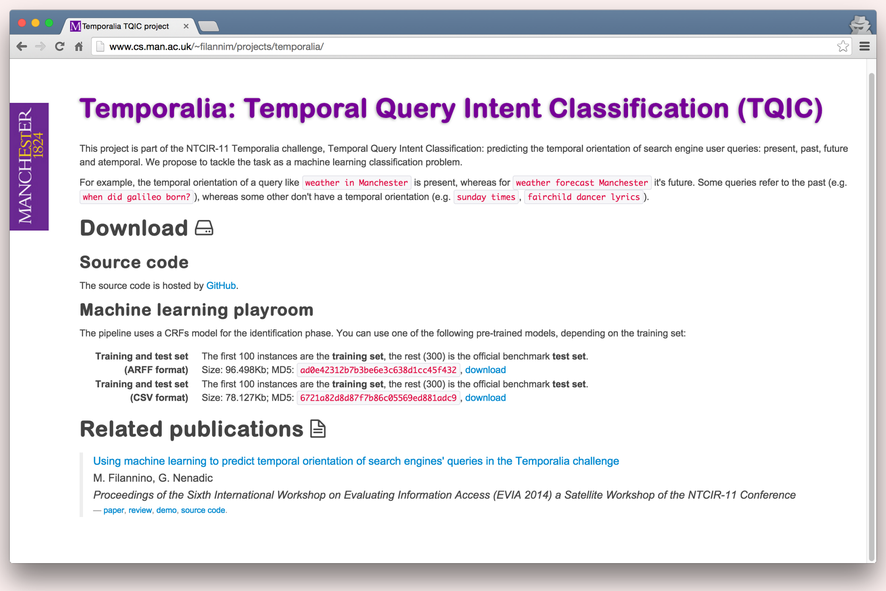
The Task

Temporal information extraction plays a crucial role in many Natural Language Processing tasks. Research in **Information Retrieval** has lead to the idea of using temporal information to improve IR systems' accuracy by guessing the temporal intent of user's queries and filtering the results accordingly.

We want to build a system which is able to predict the temporal orientation of a user's query. Temporal intents are the following: **recency, past, future and atemporal**.

Code & Resources

To aid replicability of this work, the **source code**, **datasets**, and the machine learning pre-trained **models** have been made publicly available online.



URL: http://www.cs.man.ac.uk/~filannim/projects/temporalia/

Methods: low-sparsity attributes

We approached the problem as a machine learning supervised classification task in **4 classes** (Recency, Past, Future, Atemporal).

We focussed on designing **attributes with minimal sparsity**. Some of them make use of a temporal expression extraction system, ManTIME[1], to capture queries' temporality with respect to their submission time.

The models have been trained by using the official NTCIR-11 Temporalia **training set**, which contains 100 queries along with time of submission and temporal class.

SVM with polynomial kernel (for the minimal, and intermediate runs) and Random Forests (for the full run) models have been used.

We submitted **3 runs**, each one using a different combination of attributes (see the table).

#	Input	Attribute description	Sparsity	Example Input (query/time) → attribute value	runs		
					minimal	intermediate	full
1	query	Is it a Wikipedia page title?	2	New York Times" → YES		✓	✓
2	query	Does it contain a temporal expression?	2	"june 2013 movies" → 'YES'	✓	✓	✓
3	query's	Submission's term	3	"Feb 28, 2013 GMT+0" → 'B'			✓
4	query's	Submission's trimester	4	"Aug 26, 2013 GMT+0" → 'M2'			✓
5	both	Timing	4	"Movies 2012", "Feb 28, 2013 GMT+0" → 'past'	✓	✓	✓
6	query	Most frequent trigger class	5	"peso dollar exchange rate" → 'present'	✓		✓
7	query	Wh type	5	"how did hitler die" → 'how'		✓	✓
8	query	Most frequent TempoWordNet class	5	"current stock prices" → 'present'			✓
9	query	Most frequent POS tag tense	7	"what is stop kony 2012" → 'VBZ'	✓	✓	✓
10	query	Most frequent coarse-grained POS tag	8	"kony 2012 fake" → 'N'		✓	✓
11	query	Trigger classes footprint	11	"what was I thinking lyrics" → 'past-atemporal'	✓	✓	✓
12	both	Temporal Δ between submission and query	16	"father's day 2010", "Feb 28, 2013 GMT+0" → 36.0		✓	✓
13	query	Tenses footprint	18	"when does fall start" → 'VBZ-VB'		✓	✓
14	query	Ordered TempoWordNet classes	18	"the last song" → 'past-future-present-atemporal'			✓
15	query	Most frequent fine-grained POS tag	21	"kony 2012 fake" → 'NN'		✓	✓
16	query	Coarse-grained POS tag ordered footprint	119	"when is labour day" → 'N-W-V'			✓
17	query	Fine-grained POS tag ordered footprint	202	"when is labour day" → 'NN-WRB-VBZ'			✓
18	query	Coarse-grained POS tag footprint	204	"when is labour day" → 'W-V-N-N'			✓
19	query	Fine-grained POS tag footprint	265	"when is labour day" → 'WRB-VBZ-NN-NN'			✓

Attribute list: The attributes are ordered by sparsity (number of possible attribute-values measured in the training set).

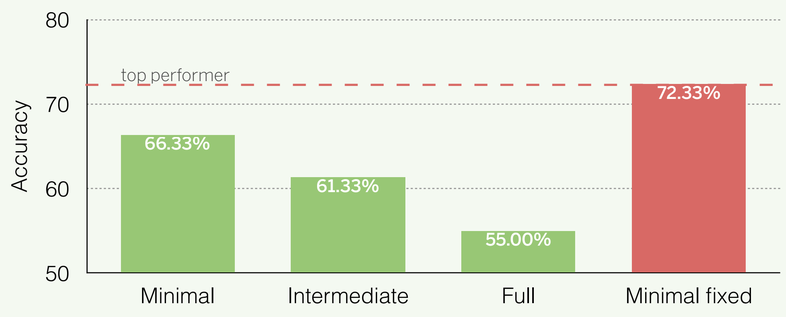
Results

The minimal run obtained the **highest accuracy** by correctly predicting the temporal orientation of 199 queries (66.33%) out of 300 in the official test set.

The model has been further improved, leading to a final accuracy of 72.33% (minimal fixed).

The minimal run ranked 5th among the best runs, and 11th out of the 17 submitted runs among all the submitted runs.

There is no **statistically significant difference** between the minimal and intermediate model, whereas the difference between minimal and full, and intermediate and full are statistically significant.



Challenge results: Results for each run are shown here. The accuracy is computed with respect to the NTCIR-11 Temporalia official benchmark test set. We corrected, a posteriori, some attributes.

The confusion matrix for the best submitted run (minimal) highlights the major sources of **classification errors**:

- future vs. recency
- atemporal vs. recency

	Classified as			
	Recency	Past	Future	Atemporal
Recency	**43**	0	21	11
Past	3	**60**	6	6
Future	38	0	**35**	2
Atemporal	6	5	3	**61**

Confusion matrix: The figures refer to the best submitted run (minimal).

[1] M. Filannino, G. Brown, and G. Nenadic. ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge. In Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 53–57, Atlanta, Georgia, USA, June 2013. ACL.

Acknowledgements: We would like to acknowledge the support of the Engineering Physical Sciences Research Council (EPRSC) and the Centre for Doctoral Training (CDT) in the form of doctoral training grant. We want to thank Gael Dias and Mohammed Hasanuzzaman for sharing with us TempoWordNet before its official public release.