TUTA1 at the NTCIR-11 Temporalia Task

Exploring Temporal Information for TQIC

Hai-Tao Yu[¶*] & Xin Kang[†‡*] & Fuji Ren[‡]

[†]Electronics and Information, Tongji University

[‡]Faculty of Engineering, The University of Tokushima

[¶]Faculty of Library, Information and Media Science, University of Tsukuba

xkang@tongji.edu.cn, kang-xin@iss.tokushima-u.ac.jp,
yuhaitao@slis.tsukuba.ac.jp, ren@is.tokushima-u.ac.jp

筑波大学
University of Tsukuba

Abstract

For NTCIR-11 Temporalia subtask Temporal Query Intent Classification (TQIC), we carefully study temporal information in the dry-run search queries, explore time gap, verb tense, lemma and named entity as temporal features, and build supervised and semi-supervised linear classifiers. We report the Precision and over Precision scores through RUN-1 to RUN-3 as well as a baseline RUN-4, compare the performance with respect to different parameter and learning algorithm configurations, and analyze the TQIC errors. We find the time gap and verb tense features with a supervised classifier are effective in separating the Past and Future queries, while the lemma and named entity feature could help predicting the Recent and Atemporal queries with a semi-supervised classifier.

Introduction

The TUTA1 group at The University of Tokushima participated in two subtasks, Temporal Query Intent Classification (TQIC) and Temporal Information Retrieval (TIR), of the new pilot task Temporal Information Access[1] (Temporalia) at NTCIR-11. The TQIC subtask focuses on the identification of user's temporal intent given the query string and submission date, across four temporal categories Past, Recent[1], Future, and Atemporal.

Challenges

1. What are effective temporal features in search queries?
2. How to explore temporal information in the background?

Temporal Feature Extraction

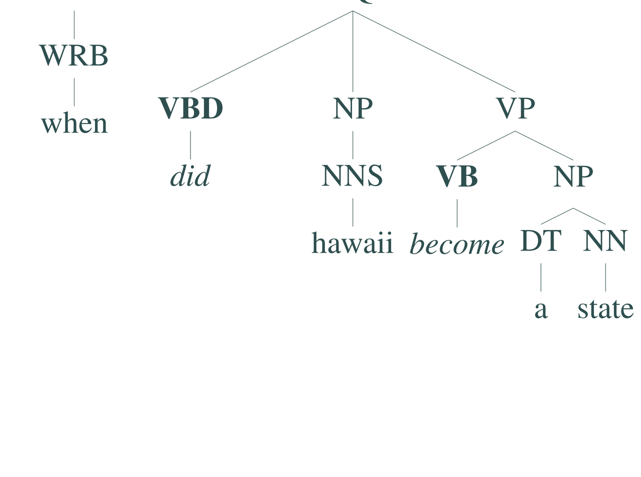
Because query strings are usually very short (4.2 words in dry-run on average), to find useful temporal features in queries and to explore the background information seem to be prominent in this subtask. AOL 500K User Session Collection[2][2] is employed to expand our knowledge of temporal features, through a semi-supervised learning model.

Class	Query String	Submit Date	Temporal Feature
Future	*June 2013* movie releases	May 28, 2013	2013-06 DIFF_future
Future	*2013 winter* weather forecast	Oct 28, 2013	2013-WI DIFF_future
Future	weather for *tomorrow*	Oct 28, 2013	P1D DIFF_future
Future	comet coming in *2013*	Oct 28, 2013	2013 DIFF_same_year
Future	comet *coming* in 2013	Oct 28, 2013	VBG UVT_VGB VGB_come
Past	when *did* hawaii *become* a state	Feb 28, 2013	VBD VB UVT_VBD VBD_do VB_become
Atemp	*New York Times*	Feb 28, 2013	ORGAN_New_York_Times
Past	*Yuri Gagarin* Cause of Death	Feb 28, 2013	PERSON_Yuri_Gagarin
Recent	*Boston Bruins* Scores	Oct 13, 2013	ORGAN_Boston_Bruins

Table 1: Extracting the time gap features.

Time Gap The ideal temporal features should indicate the gap between the intended time point in a search query and the query submission time, in which case the Temporal Query Intent Classification problem falls back to evaluating this time gap. We employ the SUTIME library in Stanford CoreNLP pipeline to recognize and normalize the temporal expressions in search queries.

Verb Tense Another important temporal feature in search query is the verb tense, which include the past tense (VBD), the singular present tense (VBZ/VBP for 3rd person/non-3rd person), the present participle tense (VBG), the past participle tense (VBN), and the base tense (VB). The verb tense features are represented by the combination of POS tags and verb lemmas. In case of multiple verbs in a query string, we use the Uppermost Verb Tense UVT_VB* to represent a user's temporal intent, in which VB* is the tense of the main predicate. Verb tense and the main predicate in a search query are obtained through Stanford POS tagger library and Stanford Parser library in Stanford CoreNLP pipeline, which follows Penn Treebank tag set for POS tagging.



Lemma and Named Entity Not all query strings contain time gap and verb tense features. We investigate the word lemmas and named entities in queries, which are still sparse but count the majority of query contents, to indicate the temporal intents. Stanford Named Entity Recognizer in Stanford CoreNLP pipeline is employed to recognize named entities in search queries.

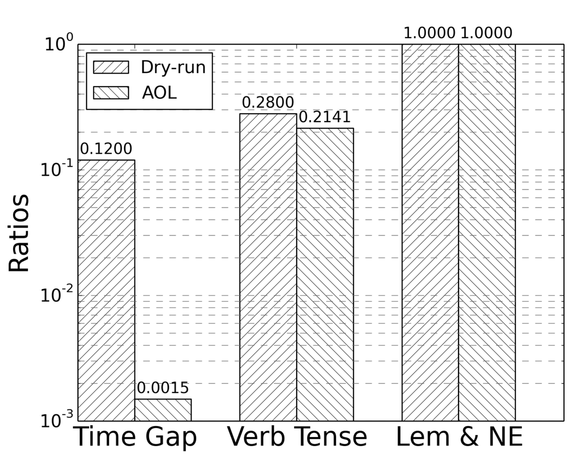


Figure 1: Temporal feature ratios.

Temporal Intent Classifiers

- Supervised classifier: Logistic Regression Classifier in scikit-learn 0.15.0.
- Semi-supervised classifier: Linear SVM Classifier in SVMlin[3].

[1]According to task description, the Recent category corresponds to the "very near past or at present time" temporal intents in search queries

[2]http://www.gregsadetsky.com/aol-data/

Experiments

Experiment Setup The supervised classifier is trained with the 80 labeled examples in dry-run dataset, while the semi-supervised classifier is trained with extra 3.1M unlabeled examples in the AOL dataset. Model parameters are selected through a 5-fold cross validation on the training dataset, based on the overall Precision. All models are tested on the formal-run dataset, which contains 300 labeled examples. RUN-1-to-3 are submitted, and RUN-4 serves as a baseline.

Run	Feature	Dataset	Classifier	Hyper Parameter
1	All temporal features	Dry-run	LGR	$C = 30, penalty = l_1$
2	All temporal features	Dry-run	LGR	$C = 300, penalty = l_1$
3	All temporal features	Dry-run, AOL	SVMlin	$A = 2, W = 0.03, U = 3, R = 0.03$
4	Lemma & named entity	Dry-run	LGR	$C = 3, penalty = l_2$

Table 2: TQIC runs.

Experiment Result TQIC results are evaluated on the formal-run dataset, based on the classification Precision for each temporal class τ

$$P(\tau) = \frac{correct(\tau)}{total(\tau)}, \tag{1}$$

and the overall Precision

$$\bar{P} = \frac{\sum_\tau correct(\tau)}{\sum_\tau total(\tau)}. \tag{2}$$

RUN-1 achieves the highest overall Precision and Precision for Future and Atemporal, while RUN-3 yields the highest Precision for Past and Recent. Results suggest that time gap and verb tense are effective in separating Past, Future, and even Atemporal, and the background information helps our semi-supervised classifier to further improve on Recent and Past.

Run	Past	Recent	Future	Atemp	Overall
1	0.8533	0.4800	**0.8533**	**0.7733**	**0.7400**
2	0.8533	0.4667^{1}	0.8267^{1}	0.7600	0.7267
3	$\mathbf{0.8667}^{1,2}$	$\mathbf{0.5867}^{1,2}$	$0.8400^{1,2}$	$0.5333^{1,2}$	0.7067
4	$0.6933^{1,2,3}$	$0.4800^{1,2,3}$	$0.8133^{1,2,3}$	$0.6400^{1,2,3}$	0.6567

Table 3: Precision scores. Wilcoxon signed-rank test with $p < 0.05$ is employed for statistical significance test: superscripts 1, 2, 3 indicate statistically significant differences to RUN-1, RUN-2, and RUN-3 respectively.

Error Analysis Recent and Atemporal seem more difficult to predict than Past and Future. In each subplot, the cell located at row i and column j corresponds to the number of observations known to be in class i while predicted as class j. For all runs, the mis-prediction of Recent to Future counts the largest number of errors, while the mis-predictions of Atemporal and Future to Recent count a significant part of classification errors. Time gap features DIFF_same_* turn to be less indicative, since they cannot suggest a useful gap. 11 mis-classifications of Future on *weather* queries indicate an over-fitting problem, since $5/6$ *weather* queries in dry-run have a Future label. 5 mis-classifications of Future on *tonight* queries, which are all labeled as Recent in formal-run, may reflect either an incorrectly learned feature or a vague boundary for Recent examples in feature space (the same query string "bruins game tonight time" of id 078 in dry-run and id 194 in formal-run, although submitted in different dates, was labeled as Future and Recent separately). The failure of understanding named entities, e.g. "belmont stakes 2013" and "voice 2013", is also responsible for some of the errors.

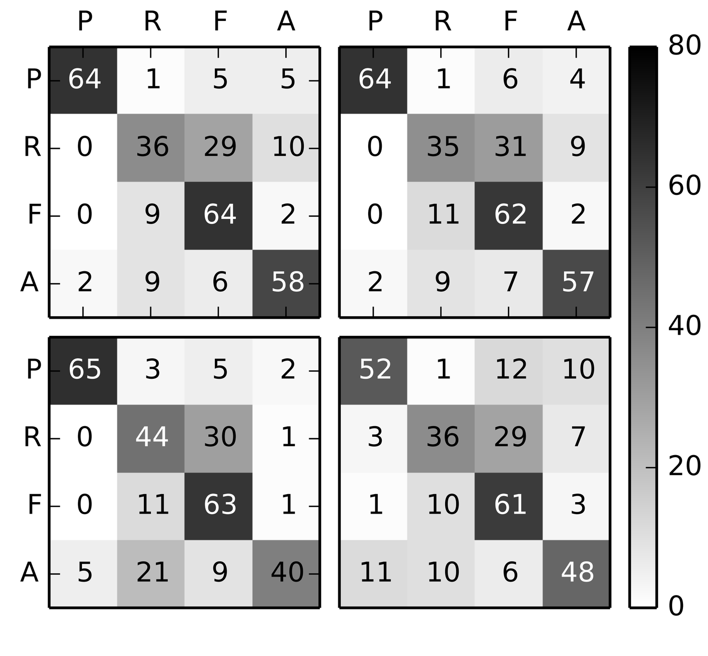


Figure 2: Confusion matrices for 4 Runs.

Conclusions

- Three temporal features were extracted for temporal information representation in search queries.
- A semi-supervised classifier was developed to expand the temporal feature on an unlabeled dataset.
- Recent-Future, Atemporal-Recent, and Future-Recent counted a big part of the mis-classification.

Forthcoming Research

Our future work will focus on investigating temporal information in lemmas and named entities. Meanwhile, methods for preventing the learning algorithms from over-fitting will also be employed.

References

[1] H. Joho, A. Jatowt, and R. Blanco. Ntcir temporalia: a test collection for temporal information access research. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 845–850. International World Wide Web Conferences Steering Committee, 2014.

[2] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *InfoScale*, volume 152, page 1. Citeseer, 2006.

[3] V. Sindhwani and S. S. Keerthi. Large scale semi-supervised linear svms. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 477–484. ACM, 2006.