Andd7 @ NTCIR-11 Temporal Information Access Task

Abhishek Shah
DA-IICT, India
201001159@daiict.ac.in

Dharak Shah
DA-IICT, India
201001069@daiict.ac.in

Prasenjit Majumder
DA-IICT, India
p_majumder@daiict.ac.in

ABSTRACT

The Andd7 team from Dhirubhai Ambani Institute of Information and Communication Technology(DA-IICT) participated in both the subtasks namely Temporal Query Intent Classification(TQIC) and Temporal Information Retrieval(TIR) of the pilot task of NTCIR-11 Temporal Information Access(Temporalia) Task [4]. This report describes different classification methods and feature sets used for classifying queries for TQIC and our approach towards building an Information Retrieval system for TIR subtask. Experimental results show that one of our system achieves the second best accuracy of all the systems submitted by different participants. Also for TIR task, we have achieved a comparative nDCG@20 which we have used for evaluation of our system.

Team Name

Andd7

Subtasks

Temporal Information Access (Temporal Query Intent Classification(TQIC),Temporal Information Retrieval(TIR))

Keywords

Temporal Intent, Information Retrieval, Query Classification

1. INTRODUCTION

The Andd7 team of DA-IICT participated in both the subtasks of pilot task-Temporalia [4]. The goal of this task is to foster research in temporal information access. Time of the document and time of the query fired plays a crucial role in determining the relevance of the document according to the information need of the user. Moreover, the temporal intent of the query is of the essence to determine the type of the documents to be retrieved. Though most of the queries are intented to ask recent information, recent analysis of this task show that there are good amount of queries searched which had intent of knowing about past incidences (e.g. History of Coca-Cola) and future predictions (e.g. release date of ios7). Thus, it is very necessary to systematize information retrieval from a temporal perspective. There are various temporal tasks organized by different consortiums like TREC Temporal Summarization Task [1] where tracking event related information was given importance and GeoCLEF [3] where it was expected to answer objectively for the questions like "When" and "Where". Temporalia task had two subtasks: Temporal Query Intent Classification (TQIC) and Temporal Information Retrieval (TIR). In TQIC, the expecation was to classify a given query in one of the temporal classes. In TIR, participants were asked to retrieve a set of documents in response to a search topic that incorporates time factor in addition to typical search topic.

2. TQIC

Temporal Query Intent Classification subtask was related to classifying a given query string to one of the temporal classes: Past, Recent, Future and Atemporal. Each of these classes have been explained in [4]. 100 queries were given for dry run and 300 queries were given for formal run of which 100 queries were used for training.

2.1 Method

Our method is mainly using well-defined classifiers like Naive Bayes Classifier, SVM(Support Vector Machine) and Decision Trees. But before classification, preprocessing step was carried out. Query class and Query issue time were parsed from the document. As time plays a crucial role to decide the nature of a query, date extractor to extract date from the query was written. Query words were stemmed using Porter Stemmer [5]. To classify a given query, the system carries out following steps:

1. Feature extractor is used to convert each input value to a features set.
2. Pairs of feature sets and labels are fed into the classification algorithm to generate a model.
3. During prediction, the same feature extractor is used to convert unseen inputs to feature sets.
4. These feature sets are then fed into the model, which generates predicted labels.

The following features have been identified:

1. Bag of words.
2. Difference of year in which query was issued and a specific year mentioned in the query. Assumption was made that any number lying between 1900 to 2100 will be considered as a year entity.
3. Number of words in the query.

4. Number of verbs in the query to specifically classify into temporal and atemporal. POS tagger of NLTK was used to tag the verbs in the query.

Different combinations of features were used from the above mentioned features. The systems were trained and tested on different combination of features with different classifiers. It was observed that classification based on feature 1 and 2 performed well than other combinations of features. As the data provided was less, more features could not be taken into account due to overfitting problem. Three systems were designed as follows:

1. For system 1, Naive Bayes Classifier was used to set the baseline.
2. For system 2, SVM (Support Vector Machine) classifier was used.
3. For system 3, Decision Tree was going to be used but it didn't perform as good as Naive Bayes or SVM. So the approach was modified and the combination of three classifiers was applied. For a given query, we have applied Naive Bayes, SVM and Decision Tree classifier to predict the class. If any two of the classifiers predict the same class then that would be the output otherwise SVM classifier predicted class would be output for that particular query. Thus, we used a combination of the results of classifiers for the System 3.

2.2 Result and Analysis

There are mainly three types of queries according to [2]:Ambiguous, Broad and Clear. A lot of ambiguous queries like "Stock Price Google" can be found in data which were not classified correctly. It becomes difficult for humans also to predict whether it is current stock price of google or stock price of google over the month or over a period of time. Apart from these, there were specific queries which were associated with certain events. To classify such queries extra information like date and time of the event are required. Some of these queries which were not classified correctly by System 3 are listed below in Table[1]:

Table 1: Queries which require extra temporal information

Query	Correct Class	Classified by System 3
season 3 game of thrones	Past	Atemporal
martin luther king day 2013	Past	Future
nba draft 2013	Future	Recent

The issue date of all these queries was 1st May, 2013 and intent cannot be determined correctly from the query because it is not possible to determine whether these events have occured before or are going to occur after the issue date without using the extra information. Moreover, the first query mentioned in the table doesn't give the true intent of the user. So user could have been searching for the star cast of the season 3 Game of Thrones, release date of the season 3 Game of Thrones, latest episode release of the season 3 Game of Thrones or may be the next episode release date. Hence it can be considered as a broad query.

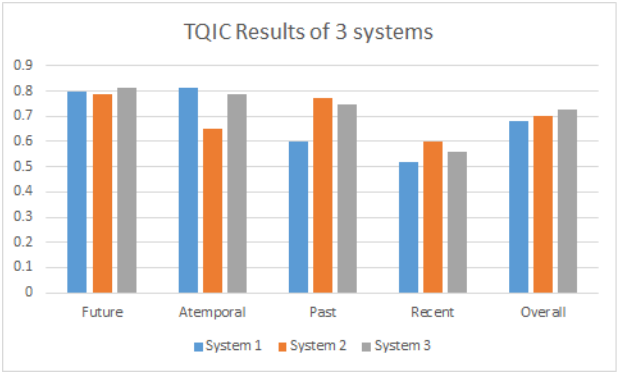

Figure 1: Accuracy of Systems for Query Classification

Dominant Keyword refers to frequently observed word in a query belonging to a specific class. In most of the cases, it was observed that presence of one or two dominant keyword decides the class of a query. List of observed some of dominant keywords for each class are as follows :

1. Future: will, forecast, shall, upcoming, next
2. Past: history, was, were, past, biography
3. Recent: is, today, now, current, live
4. Atemporal: what, how

As it can be observed from the Figure 1, accuracy of classification of Future and Past classes is high compared to Recent and Atemporal class. We observed that overlapping of query words from Recent and Atemporal classes were high, which resulted in mis-classification of query and hence Recent class has suffered in results comparatively decreasing the overall result. System 3 showed the second best result among all the submissions with **72.6%** accuracy. Analysis of the the class-wise accuracy of all the 3 systems gives a better insight. For Recent and Past classes, System 2 performed better than System 3. For many queries of Recent and Past classes, SVM classifier gives the best result. For Atemporal class, System 1 performed better than System 2. Naive Bayes classifier performed significantly better than SVM. For Future class, System 1 and System 2 has performed significantly better than other classes. So, as we designed System 3 with combination of all classifiers we overshadow poor performance of each class and try to get the best result for each class. Thus, System 3 performed better overall.

3. TIR

In this subtask of Temporal Information Access, it was expected to retrieve documents considering the temporal and topical relevance of the document to the user need.

3.1 Data

For this task, we indexed a 20 GB uncompressed document corpus [7] called "LivingKnowledge" which contained annotated news articles and blogs collected from about 1500 sources using Apache Lucene [8]. It had around 3.8M articles ranging from dates May 2011 to March 2013. It had different fields of document id, date of the document, hostname,

host url, title of the document, content of the document etc. The temporal data and named entities were tagged in the content section of the document sentence-wise. For example, considering a sentence from one of the documents of 2012.
$< Tval = "20120530" > 30May < /T >$
All such date entities were indexed after preprocessing in the content separately document wise. The 'val' attributes of tag T were taken for the dates and were used in the same format. At some places only month or year was mentioned. Published day of the document and published month and day of the document respectively were added to such 'val' attributes to make such entities having all day, month and year. For example,
$< Tval = "2012" > 2012 < /T >$
If this tag is mentioned in document date whose published date is 30th May, 2012, then we would change the 'val' attribute to '20120530'. We were given 50 topics for formal runs and a typical topic had a title and a subtopic for each type of classes: Atemporal, Recent, Past and Future. Also it had time and date at which query was fired.

3.2 Method

User needs in this case were topics as described above. Each topic had a title, description and subtopics of classes: Recent, Past, Future and Atemporal and date and time when that particular topic is searched. As it was not allowed to use the classes of the topics, certain words were identified from TQIC task's queries which were provided during dry runs and were very common in definite temporal class of queries. We also included synonyms of some of the words and name of the classes itself. This classification was done as one of our systems employ method which is class dependent and has been explained later in this section. Main three temporal classes and words that were used in association to that are as follows:

1. Future: future, forecast, will, would, should, shall, next
2. Recent: recent, present, current, latest, live
3. Past: history, were, was, past, origin, did, start, been

We used Apache Lucene Framework for making our systems and we designed two types of queries:

1. A query containing title of the topic to be searched in the document title field and subtopic to be searched in the content field.
2. A query having only subtopic to be searched in both title and content field of the document.

The title was filtered by removing all the unnecessary words and keeping only nouns and verbs in the title and adjectives present in each of the subtopic were also considered in the query. Tagger of Stanford CoreNLP package [6] was used for the above purpose. The adjectives were used because there were some titles like "Waterborne diseases in Africa". Here, 'waterborne' becomes a very important word otherwise a lot of non-relevant documents related to disease in Africa would have been ranked higher. 3 systems were developed as follows:

1. System 1 was developed where in topical search was done and query contained only the subtopic which was searched against the title field of the document as well as content field of the document.
2. In System 2, only topical search was done and query contained the title of the topic which was searched against the title field of the document and subtopic was searched against the content field of the document.
3. In System 3, topical search and temporal re-ranking were used and topical search was done in the same manner as in System 1.

BooleanQuery was prepared to search mentioned attributes of topic and the queries were parsed using Standard Analyzer of Apache Lucene [8]. In System 2, the filter for the title as explained above was also used.
In system 3 along with topical retrieval of Apache Lucene, a temporal based re-ranking scores were also added for different query classes for each query. The temporal scores for different class of queries were calculated as shown below:

1. Future query: temporal score of a document= $\frac{(\mu-d)-min((\mu-d))}{max((\mu-d))-min((\mu-d))}$
2. Recent query: temporal score of a document = $\frac{(|\mu-d|)-min((|\mu-d|))}{max((|\mu-d|))-min((|\mu-d|))}$
3. Past query: temporal score of a document= $\frac{(d-\mu)-min((d-\mu))}{max((d-\mu))-min((d-\mu))}$

where μ = mean of dates in the document
d = date on which document was published
$max(y)$ = defines maximum y for all the documents retrieved for that query
$min(y)$ = defines minimum y for all the documents retrieved for that query
final_score for System 3 =
$0.5 \times topical\_score + 0.5 \times temporal\_score$
If there are no temporal tags only topical_score would be considered with complete weightage unlike equal weightage to both the scores. The motive behind this ranking was that if the dates which have been mentioned in the document are near the document date, the document is talking about the things which have happened in recent times. If the dates are quite older than document date it is talking in the past sense and similarly if they are later than document date then it is talking about future. Document fields like document date, document title, document content were used in different systems as explained above. In the given topic, we had used title in two of our systems, subtopic for each type were used to enhance the topical search.

3.3 Result and Analysis

Our word based classifier of topics could classify queries with accuracy of 86%. The queries which were not correctly tagged were by default tagged Atemporal. Moreover, around 2.6M documents had at least one temporal tag in their content field and hence those documents carry important information for temporal relevance. Of all the documents in the pool of relevance assessment statistics of the documents having at least one temporal tag is shown in Table[2]. Here note that same document could be marked relevant for one query and not relevant for other. Thus, it can be observed that pool chosen has significant number of relevant and partially relevant documents as compared to number of non-relevant documents. The performance of the 3 systems are given in Table[3].

Table 2: No. of documents having atleast one temporal tag according to relevance

Relevance	No. of documents
Not relevant	10570
Partially relevant	6864
Relevant	5499

Table 3: nDCG@20 of 3 systems for different query class

	System 1	System 2	System 3
Atemporal	**0.3853**	0.3841	**0.3853**
Future	0.3203	0.3280	**0.4030**
Past	0.2739	0.2815	**0.3327**
Recency	**0.4888**	0.4884	0.4171
Overall	0.3671	0.3705	**0.3845**

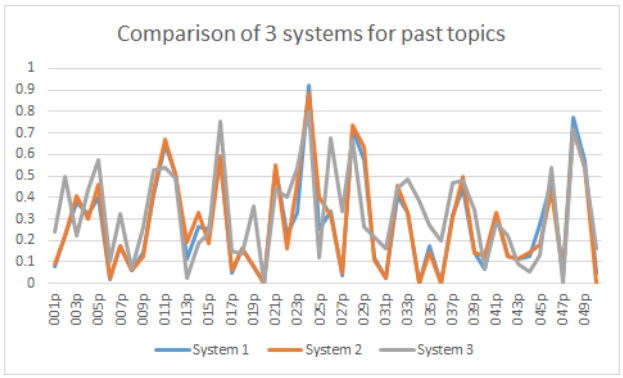

Figure 2: Comparison of 3 systems for past topics

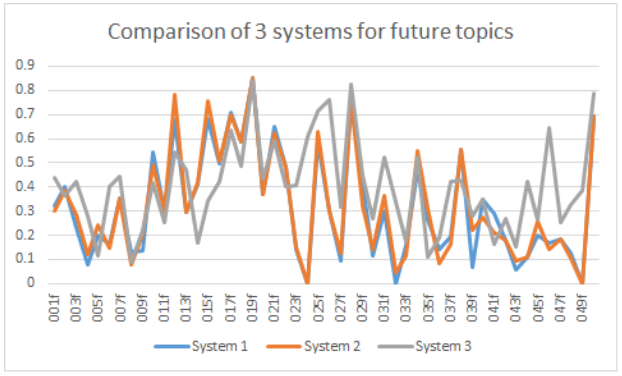

Figure 3: Comparison of 3 systems for future topics

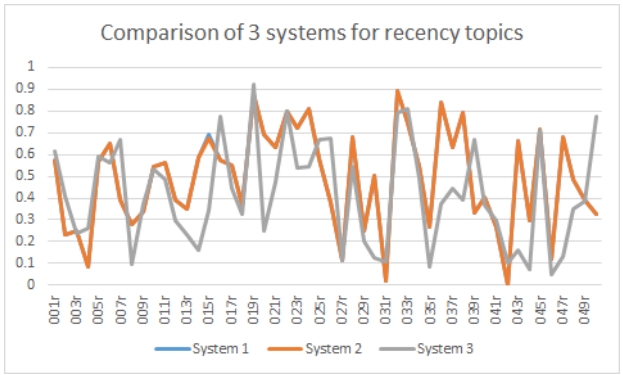

Figure 4: Comparison of 3 systems for recency topics

It can be observed that performance considerably increases for the Past and Future queries for System 3 due to the additional re-ranking according to the temporal relevance. As mentioned in the above formula, we also tried to substitute the query date instead of document date but the system wouldn't perform well as most of the query dates would prove the documents to be of the past nature. For recent class of queries, the system without temporal re-ranking gives better performance. That means for atemporal and recent queries, topical search can provide good results. Comparison of all three systems for future, present and past set of topics have been shown in Figure 2, Figure 3 and Figure 4. System 3 i.e. system with temporal re-ranking performs better overall. There is hardly any difference in the performance of System 1 and System 2. For past topics, it can be observed that all the three systems show very similar pattern which show that certain topics are responsible for bringing down the score of all the systems. Examples of past topics which were correctly classified but for which all systems have performed poorly

1. What were the early treatment options for diabetes?
2. What major earthquakes occurred in Japan in the past before the 2011 Tohoku earthquake?

For future topics, also there are only few queries where we see the difference between the performances of the systems like ẄHat will the implications be for corporate entities if the First Amendment has been changed?T̈here were certain subtopics for which all the three systems fail like:

1. How do people describe the personality of JK Rowling?
2. What were the past actions or suggestions for solving the problem of the ageing population in the world?

These queries were accurately classified but still none of the systems could rank a relevant document in top 20 retrieved documents. The topical search of the query needs to be improved drastically and also relevant temporal information in the content needs to be identified considering the context around the temporal tags which would have improved the performance because all the temporal information is not necessarily relevant.

4. CONCLUSIONS

In this paper, we proposed methods for Temporal Query Intent Classification and Temporal Information Retrieval subtasks for the pilot task of Temporal Information Access in the NTCIR-11. In Temporal Query Intent Classification, this report shows performance of different classifiers using some basic features. Due to less training data, increase the accuracy for determining Recent Class Queries and Atemporal Queries became tough due to which our results suffered. Some queries were ambiguous to classify into one of the above mentioned classes. In TIR task, we used dates provided in the document's content and compared the mean of such content dates and compared it with published document date for determining the nature of the document. All the dates in the given document were considered as relevant temporal information for the particular subtopic class. By ranking those documents higher than other documents for that class of queries considering the temporal score and topical score, we achieve better performance. Results showed the effectiveness of the systems in terms of exploiting temporal nature of the documents and query. But it can be improved by identifying the temporal nature relevant to the subtopic and better results can be achieved.

5. ACKNOWLEDGEMENTS

We sincerely thanks Mr. Nitin Ramrakhiyani for his suggestions in research work and valuable insights and reviews for the paper.

6. REFERENCES

[1] J. Aslam, F. Diaz, M. Ekstrand-Abueg, P. Virgi, and T. Sakai. TREC 2013 Temporal Summarization. 2013.

[2] R. N. T. Campos. *Disambiguating Implicit Temporal Queries for Temporal Information Retrieval Applications*. PhD thesis, Universidade do Porto, 2013.

[3] F. Gey, R. Larson, M. Sanderson, H. Joho, P. Clough, and V. Petras. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. 2013.

[4] H. Joho, A. Jatowt, R. Blanco, H. Naka, and S. Yamamoto. Overview of NTCIR-11 Temporal Information Access(Temporalia) Task. 2014.

[5] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[6] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP Natural Language Processing Toolkit.

[7] M. Matthews, P. Tolchinsky, R. Blanco, J. Atserias, P. Mika, and H. Zaragoza. Searching Through Time in the New York Times. 2010.

[8] M. McCandless, E. Hatcher, and O. Gospodnetic. Lucene in Action: Covers Apache Lucene 3.0. 2010.