Entity Summarization of News Articles

Gianluca Demartini*†
L3S Research Center
Appelstrasse 9a
30167 Hannover, Germany
demartini@L3S.de

Malik Muhammad Saad Missen*
IRIT
Toulouse, France
missen@irit.fr

Roi Blanco, Hugo Zaragoza
Yahoo! Research
Diagonal 177
08018 Barcelona, Spain
{roi,hugoz}@yahoo-inc.com

ABSTRACT

In this paper we study the problem of entity retrieval for news applications and the importance of the news trail history (i.e. past related articles) to determine the relevant entities in current articles. We construct a novel entity-labeled corpus with temporal information out of the TREC 2004 Novelty collection. We develop and evaluate several features, and show that an article's history can be exploited to improve its summarization.

Categories and Subject Descriptors:
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms:
Algorithms, Measurement, Experimentation

Keywords:
Entity Summarization, Time-aware Search

1. INTRODUCTION

Entity retrieval is becoming a major area of interest in IR research and it is quickly being adopted in commercial applications. One of the promising areas applying entity retrieval models in the commercial world is news search. News retrieval has also been the focus of much attention in the IR research community, but to our knowledge there have been no entity ranking tasks defined for news.

Consider the following user scenario: a user types a query (or topic) into a news search engine and obtains a list of relevant results, ordered by time. Furthermore, the user subscribes to this query so in the future she will continue to receive the latest news on this query. We are interested in entity ranking tasks related to this user scenario. For instance, standard entity ranking could be used to show the most interesting entities *for the query*. In practice, the temporal dimension is not needed here. However, if the user is observing a current document, we may want to show the most relevant entities of the document for her query taking into account features extracted from previous documents. This prompts the Entity Summarization (ES) task definition: given a query, a relevant document and possibly a set of previous related documents (the *history* of the document), retrieve a set of entities that best summarize the document. This is a newly defined task that can be useful, for example, in vertical search for presenting the user more than just a ranked list of documents.

*Work performed while intern at Yahoo! Research.

†This work is partially supported by the EU Large Scale Integrated Project LivingKnowledge (contract no. 231126).

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

2. TIME-AWARE ENTITY SUMMARIZATION

More formally, we define a "news thread" relevant to a query as the list of relevant documents $D = [d_1 \dots d_n]$ chronologically ordered. Then, given a document d_i we define its history as the list of relevant documents $H = [d_1 \dots d_{i-1}]$ chronologically ordered pre-dating the document d_i. Given an entity e, we note as $d_{e,1}$ the first document in which the entity occurred in the news thread. Note that such a document is not necessarily the first document in D as entities may appear only in subsequent documents. Moreover, we note as $d_{e,-1}$ as the last document in H which contains e.

For addressing this task, we propose features both from the local document as well as from H. The first feature we consider is the frequency of an entity e in a document d, noted $F(e,d)$. In the following we will use this feature as our baseline. It is possible to consider if an entity appears as a subject of a sentence as this is generally the person or thing carrying out an action (after running a dependency parsing over the sentence collection). Hence, we define the $F_{subj}(e,d)$ as the number of times an entity e appears as subject of a sentence in the document d.

Additionally, we propose two position-based features that take into account where in document d an entity e appears. Let $FirstSenLen(e,d)$ be the length of the first sentence where e appears in document d and $FirstSenPos(e,d)$ be the position of the first sentence where e appears in d (e.g, the fourth sentence in the document).

We now introduce a number of features that take into consideration the document history H. Let $F(e,H)$ be the frequency (i.e., the number of times it appears) of the entity e in the history H. Instead of counting each entity occurrence a simpler variation considers the number of documents in which the entity e has appeared so far. We thus define $DF(e,H)$ as the document frequency of e in H.

Furthermore, it is possible to examine single documents from the past to extract more features; we then define $F(e,d_{e,-1})$ as the frequency of entity e in the previous document where the entity appeared and $F(e,d_{e,1})$ as the frequency of entity e in the first document where the entity appeared.

We can also compute $CoOcc(e,H)$, the number of other entities with which the entity co-occurred in a sentence in the set of past documents H.

3. EXPERIMENTAL EVALUATION

We selected the 25 event topics of the latest TREC Novelty collection (2004) consisting of news articles. We annotated the documents associated with those topics using state of the art NLP tools[1] in order to extract entities of type person, location, organization, and product based on WSJ annotations. The system detected 7481 entity occurrences in the collection: 26% persons, 10% locations, 57% organizations, and 7% products. Human judges assessed the relevance of the entities in each document with respect to the topic grading each entity on the 3-points scale: Relevant, Related, Not Relevant. An additional category was used, i.e., 'Not an entity', to mark entities which had been wrongly annotated by the NLP tool. A total of 21213 entity-document-topic judgements were obtained in the collection[2].

We compare the effectiveness of different features and some feature combinations using several performance metrics. We report values for Precision@3 (P@3), Precision@5 (P@5), and Mean Average Precision (MAP) considering Related entities as non-relevant and using tie-aware metrics [2].

Feature	P@3	P@5	MAP
All Ties	.34	.34	.42
Individual Features (Local and History)			
F(e,d)	**.65**	**.56**	**.60**
FirstSenLen	.37	.36	.45
FirstSenPos	.31	.31	.43
F_{subj}	.49	.44	.50
$F(e, d_{e,1})$	.58	.53	.56
$F(e, d_{e,-1})$	.64	.56	.62*
$DF(e, H)$	.63	.57*	.65**
$F(e, H)$	**.66**	**.59****	**.66****
$CoOcc(e, H)$	.62	.57	.65**
Features combined with F(e,d)			
$FirstSenLen$	.65	.57*	.62**
$FirstSenPos$	**.67****	**.58***	**.62****
F_{subj}	.65	.56	.61
$F(e, d_{e,1})$	.65	.57**	.61**
$F(e, d_{e,-1})$	.68**†	.60**	.65**
$F(e, H)$	**.70****††	**.62****††	**.68****††
$CoOcc(e, H)$	.68**††	.61**††	.67**††
$DF(e, H)$	.69**††	.61**††	.68**††

Table 1: Effectiveness of individual features and of features when combined with $F(e, d)$. Bold values indicate the best performing runs. * (**) indicates statistical significance w.r.t. F(e,d) and †(††) w.r.t. F(e,H) with paired t-test p<0.05(0.01).

Individual Features. The upper part of Table 1 shows effectiveness values obtained when ranking entities in a document according to individual features. For comparison, a feature that assigns the same value to each entity would obtain a MAP value of 0.42. The feature $F(e, d)$ obtains the best MAP value (0.60) among features from the local article. In general, history features perform better than local features and the highest performance is obtained by ranking entities according to their frequency in the past documents. Interestingly, when identifying relevant entities for a document, the frequency of the entity in the previous document in the story $F(e, d_{e,-1})$ is a better evidence than the frequency in the current document. This may be an indication of how people read news: some entities become relevant to readers after repeated occurrences. If an entity appears also in the previous documents it is more likely to be relevant.

Given these results we conclude that the evidence from the past is very important for ranking entities appearing in a document. We expect effectiveness of methods that exploit the past to improve as the size of H grows. That is, the more history is available the better we can rank entities for the current news. For $|H| \approx 20$ the average effectiveness of $F(e, H)$ grows together with $|H|$ up to values of 0.7 MAP.

Combined Features. So far we have presented different features for ranking entities that appear in a document. Combining them in an appropriate manner yields a better ranking of entities; however, because the probability distribution of relevance given a feature is different among features we need a way for combining them. The following experiments rank entities in a document according to a score obtained after combining several features together. We consider linear combination of features (transformed with a function as explained in [1]).

Let the score for an entity e and a vector $\vec{f}$ of n features be $score(e, \vec{f}) = \sum_{i=1}^{n} w_i g(f_i, \theta_i)$, where w_i is the weight of each feature and g is a transformation function for the feature f_i using a given parameter θ_i. In this paper we employ a transformation function of the form: $g(x, \theta) = \frac{x}{x+\theta}$ as suggested in [1], where x is the feature to transform and θ is a parameter. We also tried a linear transformation but it did not perform as well (more complex non-linear transformations could also be explored). In order to combine features we then need to find a parameter θ_i for the function g and a weight w_i for each feature f_i. We tested two and three features combinations, where the variables θ_i, and the combination weights w_i have been tuned with 2-fold cross validation of 25 topics training to optimize MAP. In order to find the best values we used a optimization algorithm that performs a greedy search over the parameter space [3].

Combining $F(e, d)$ with another feature is able to outperform the baseline for some range of the weight w that can be learned on a training set. The best effectiveness is obtained when combining $F(e, d)$ and $F(e, H)$ obtaining an improvement of 13% in terms of average precision. Other features, when combined with the baseline, also obtain high improvements performing as good as the combination with $F(e, H)$ ($CoOcc(e, H)$ having 12% and $DF(e, H)$ having 13% improvement in terms of MAP).

As future work, besides testing our features on different time-aware document collections, we aim at adopting machine learning techniques to combine the proposed features.

[1] `http://sourceforge.net/projects/supersensetag/`

[2] The evaluation collection we have created is available for download at: `http://www.l3s.de/~demartini/deert/`

4. REFERENCES

[1] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *SIGIR '05*, USA. ACM.

[2] F. McSherry and M. Najork. Computing information retrieval performance measures efficiently in the presence of tied scores. In *ECIR*, 2008.

[3] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, (4), 2009.