A Weakly-Supervised Detection of Entity Central Documents in a Stream

Ludovic Bonnefoy
University of Avignon
CERI-LIA / iSmart
ludovic.bonnefoy@alumni.univ-avignon.fr

Vincent Bouvier
Aix-Marseille University
LSIS CNRS / Kware
vincent.bouvier@lsis.org

Patrice Bellot
Aix-Marseille University
LSIS CNRS
patrice.bellot@lsis.org

ABSTRACT

Filtering a time-ordered corpus for documents that are highly relevant to an entity is a task receiving more and more attention over the years. One application is to reduce the delay between the moment an information about an entity is being first observed and the moment the entity entry in a knowledge base is being updated. Current state-of-the-art approaches are highly supervised and require training examples for each entity monitored. We propose an approach which does not require new training data when processing a new entity. To capture intrinsic characteristics of highly relevant documents our approach relies on three types of features: document centric features, entity profile related features and time features. Evaluated within the framework of the "Knowledge Base Acceleration" track at TREC 2012, it outperforms current state-of-the-art approaches.

Categories and Subject Descriptors

H.3.1 [**Information Storagei and Retrieval**]: Information filtering

General Terms

Experimentation

Keywords

data stream, entity linking, information filtering, kba, named entity disambiguation, time

1. INTRODUCTION

Information about popular entities on knowledge bases (KB) like Wikipedia are almost updated in real-time. According to [5] the median time-lag between the first appearance of a new information about an entity and its publication on Wikipedia is 356 days. This delay may however be reduced if relevant documents are automatically found as soon as they are published and then presented to the contributors.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

A two step process is then required: entity disambiguation (resolving to which entity in the KB a name in a document is referring to) and the evaluation of the importance of the information contained in the document with regards to the given entity.

State-of-the-art approaches are highly supervised and require a training set for each entity. A real-world system, however, must work without additional examples for new entities whatever its type, its degree of popularity, and its evolution through time are.

We propose to automatically assess relevance of a document from a data stream with regards to a given entity without requiring additional data. This approach relies on three complementary types of features to capture characteristics of relevant documents: time-related, document and entities centric features.

We evaluate it in the framework provided by the "Knowledge Base Acceleration" task at TREC 2012 and it performs better than existing approaches. Moreover, we draw preliminary conclusions about characteristics of highly relevant documents, independently of which entity is monitored.

2. RELATED WORK

Thanks to the *Text Analysis Conference (TAC)*[1] with the "Knowledge Base Population" task [7], a lot of work have been done on named entity disambiguation. Best approaches rely on similarity between a mention's context and a candidate knowledge base entry [1], name string matching [10], query expansion [6], topic modeling [14] or coreference resolution [3]. Most of theses approaches rely on computationally expensive features to evaluate the importance of the information in the documents.

Recently, named entity disambiguation in data stream have emerged relying on data from Twitter. [9] for instance followed the evolution of big and short terms events, like natural disasters, in real-time. Unfortunately, because of the characteristics of Twitter around which such approaches have been built (very short texts, hashtags, user profiles, etc. [4]), methods cannot be transposed to our problem.

A decade ago, a TREC task called "Filtering" [11] had the following definition: finding documents relevant to a query in a stream of data. Several effective approaches were inspired by information retrieval techniques to score documents (Okapi [12], Rocchio [13], ...) with the use of a learned threshold to filter out non relevant documents [15]. Most successful approaches rely on machine learning with exten-

[1]http://www.nist.gov/tac/2012/KBP/index.html

sive use of SVMs with words as features [2]. Systems were, however, in an ideal scenario: after each decision, they were notified of the annotator's label, allowing them to reevaluate their models and do not propagate their mistakes.

In 2012, a new TREC task called "Knowledge Base Acceleration" (*KBA*) [5] started with a similar definition: filtering a time-ordered corpus for documents that are highly relevant to a predefined list of 29 entities from Wikipedia and chosen for their ambiguity. The main differences between the two tasks are: a collection 10 times larger, various types of documents, finer-grained time unit.

As for the *Filtering* task, the best performing approach at *KBA* 2012 is highly supervised: one classifier (SVM) by entity tracked with "binary features, representing whether or not a term was present in a document, regardless of its frequency" [8]. In this setup, training data have to be provided for each new entity "followed" and even for an already monitored entity, new training examples are required to prevent a performances decrease due to concept drift through time.

Our approach, while achieving better results, differs from this work in that rather than trying to determine characteristics of a relevant document for a given entity, we focus on features of relevant documents in general.

3. LEARNING TO DETECT HIGHLY RELEVANT DOCUMENTS IN A STREAM

Let us consider a stream of documents from various types (news, forum, blogs, web pages, etc.). We want to monitor this stream, detect documents referring to a given entity e and then select highly relevant documents.

We tackle this challenge as a binary classification problem: highly relevant documents vs. mentioning, non mentioning and spams. Numerous works on text and genre classification proposed a wide range of features to associate a class to a document. Generally, the wider the set of features is, the better the results are (some approaches rely on thousands features). A good classification approach, however, have to find the good trade-off between good results (depending of the amount of features) and runtime. In our scenario, a fast approach is required to deal with the large amounts of incoming documents. Moreover, we do not want to use entity specific features. We rely on a set of 35 computationally inexpensive features falling in three categories: time related, document centric, entity's profile related.

3.1 Document centric features

The first source of information is the content of the document itself. To design effective features we looked at summarization works as they look for evidence to determine the topic(s) of a document. The frequency of the tracked entity is a first indicator about the relevance of the document. We also count the number of sentences mentionning the entity as how well distributed the occurrences are seems important. Titles, as well as beginning or ending of documents (especially news) have been shown to carry a lot of information about the main topic of a document so we count the number of mention of the entity in these different parts. We compute these features using strict name matching. How much the document is focused on a single or a few topics is important to and may be reflected by its entropy. The type of the document seems important too as different types of documents may not follow the same "rules". The complete list of document centric features is presented in Table 1.

$TF(e,d)$	Term frequency of the entity e in d
$hasTitle(d)$	Does the document have a title?
$TF(e,t_d)$	Term frequency of e in the title
$TF_{10\%}(e,d)$	Term frequency of e for each 10% part of d
$TF_{20\%}(e,d)$	Term frequency of e for each 20% part of d
$C(sent,e,d)$	Count of sentences mentioning e
$entropy(d)$	Entropy of document d
$length(d)$	Count of words in d
$is(social,d)$	Is d a blog entry or a forum thread?
$is(news,d)$	Is d a news?

Table 1: Document centric features. TFs are normalized by the size of the document.

3.2 Entity related features

Previous features look at the nature of the document itself, independently of the entity considered. But how the document fits to what we know about the entity seems important too. We suppose that one representative document about the entity is provided to (or retrieved by) the system. This document will be called the "source document". In our experiments the source document of an entity is its Wikipedia page. A candidate document is judged on how much related entities appear in it and how similar the document is to the source document. We apply a named entity recogniser [2] to extract a first set of related entities; a second set is created by filtering out, from the first one, entities not embedded in a link. Similarity between documents is measure with the cosine similarity with tf-idf weigths based unigrams or bigrams and without prepossessing or smoothing. Features are listed in Table 2.

$SIM_{1g}(d,sd)$	Cosine similarity between d and the source document sd based on unigrams
$SIM_{2g}(d,sd)$	Cosine similarity with bigrams
$TF(re,d)$	Term frequency of related entities in d
$TF(reL,d)$	Term frequency of related entities (embedded in links) in d

Table 2: Entity related features. If applicable, features are normalized by the size of the document

3.3 Time features

Exploring the impact of time related features is one of the most interesting characteristics of the studied scenario. Our hypothesis is that if something important about an entity happens, in (and for) a short time-span, the number of documents mentioning the entity may suddenly grow. We designed a set of features based on this intuition, listed in Table 3.

4. EXPERIMENTS

4.1 TREC KBA 2012 Framework

We evaluate the proposed approach within the framework of the KBA track which provided evaluation on a corpus of 500 million documents ($\approx$ 9Tb). Documents are either blog or forum posts (*social*), *news* or *web pages*. Documents were

[2]StanfordNER

$TF(e,d).IDF(e,1h)$	Term frequency in d and inverse document frequency for an hour
$DF(e,1day)$	Number of documents with e this day
$DF(e,7d)$	Number of documents with e in 7 days
$Var(DF(e,7d))$	Variance of the DF in 7 days
$TF(e,7d)$	Term frequency of e in 7 days
$TF(e,title,7d)$	TF of e in titles in 7 days

Table 3: Time features.

crawled from October 2011 to April 2012 and to each document is associated a time-stamp corresponding to its date of crawl. For training purpose, the corpus have been split with documents from October to December as examples (with only *social* and *news* documents) and the remainder for the evaluation. The 29 entities correspond to persons and organizations that exist in Wikipedia. Two evaluations were provided: finding documents mentionning an entity and finding centrally relevant documents defined as "documents containing information which may worth a knowledge base update". In this work we focused on finding centrally relevant documents as it is the harder task. Participants must provide one ranked list of documents for each entity. The official metric was the F-measure (harmonic mean between precision and recall)[3].

4.2 Classification

We proposed a set of features exploring characteristics of centrally relevant documents. A random forest classifier (RF) is composed of several decision trees, each one using a subset of the features and the examples. For a test document, the class receiving the most votes (one tree = one vote) is associated to it. We report results obtained with RF as they are among the best we got and give some insight on what features do well.

We decide to use two classifiers in cascade to evaluate documents: one for filtering out non mentioning documents and the other to dissociate poorly relevant documents from centrally relevant ones. Each classifier relies on all the features presented.

4.3 Results

Table 4 shows results of our approach against the best system at KBA 2012, the median system among participants and the mean. Our approach, listed as "All", achieves state-of-the-art results: with a score of .382, it performs better than the best 2012 TREC KBA system (+6%) and far better than median system (+32%) and mean (+73%).

Run	F-measure	Run	F-measure
Our approach	.382	Median KBA	.289
Best KBA	.359	Mean KBA	.220

Table 4: F-measure of our approach against best, median and mean at KBA 2012.

In addition to its good performances, we claim that the huge advantage of this approach is that it does not require additional training data for new entities. The best KBA system used one SVM classifier by entity [8] with words as features and requires training examples for each entity tested. We evaluate how well our system does without specific training data for a monitored entity by removing from the training set, examples associated to it. Under this configuration, our system gets an F-measure of .361 (reported as "1vsAll"). This result, still above the one of the best KBA system, shows that our approach succeed to capture intrinsic characteristics of centrally relevant documents, independently of the entity evaluated. The decrease may be explained by the non uniform amounts of examples associated to each entities (see [5]): setting aside training data associated to some entities can dramatically decrease the number of examples to train the models.

[3]Evaluation is made for different subsets of the result lists and the best score is selected. The cutoff is the same for all entities.

Run	F-measure	Run	F-measure
1vsAll	.361	cross3	.354
cross10	.355	cross2	.339
cross5	.350		

Table 5: F-measure using different folds. Scores are the mean of the scores of 10 runs.

To test the robustness of our approach, with regard to the amount of training data, we evaluate several configurations by partitioning the set entities in n sets and evaluated each part with the training data associated to the others $n-1$ ones (as for cross-validation). To reduce variability, for each $n \in \{10, 5, 3, 2\}$, 10 runs are made and averaged results are listed Table 5 (cross$n$). As expected, the smaller the set of training examples is, the lower the performances are. The results are however still high and above the median and mean: for $n = 10$ it would be ranked 2nd and in 3rd position for $n \in 5, 3, 2$.

Random forests provide information on how well features helps to separate classes and give insight on which ones help to characterize centrally relevant documents about an entity in a stream. The mean decrease Gini score associated by a random forest to a feature is an indicator of how much this feature helps to separate documents from different classes in the trees. Figure 1 reports these scores. Not surprisingly, top 3 criterion are profile related features: the similarity of a document with the source document (the Wikipedia page of the entity) and the presence of related entities seems to be good predictors of the importance of an information in a document. In the top 10, the three types of features are represented (time, document centric and profile), showing that all the sources are complementary. Beyond this rank, features seem not to be quite useful to distinguish classes of documents. We re-evaluated our system using only these top 10 features and results confirmed this observation : .355 versus .361 with all of them. More surprisingly, the title seems useless: its presence in a document does not seems to influence the output of the classification and the presence of the entity in it is the less discriminative feature. This result is not in line with research on summarization which showed that titles are good indicators of the content of a document. The presence of an entity in a title with regards to relevance and the correlation is indeed very weak: only 53% of documents with the entity in a title are relevant. Positions of mentions in the document are not discriminative either. Finally, knowing the nature of a document does not help to take decisions even if our approach gets better results

for documents from *news* and *social* categories than for *web pages* (probably because there is no document of this type in training data).

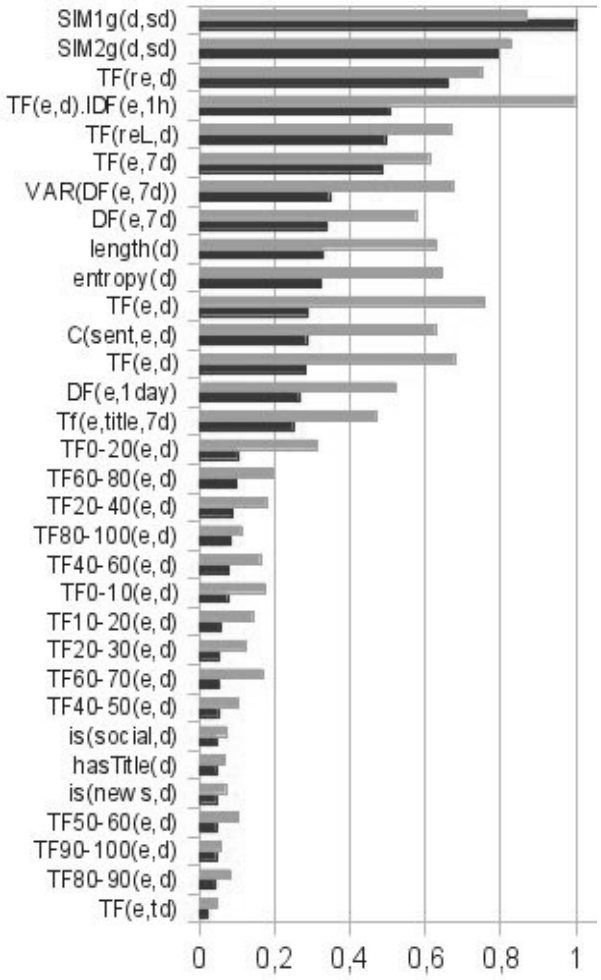

Figure 1: Mean decrease Gini score for non-mentioning/mentioning (black) and mentioning/centrally relevant (grey)

5. CONCLUSIONS AND FUTURE WORK

We propose to detect documents containing highly relevant information about an entity in a stream of data by capturing intrinsic characteristics of these documents. This weakly supervised approach rely on time, document centric and entity profile related features. It showed state-of-the-art performances and do not require additional training data to work with new entities. Moreover, it is robust enough to still be competitive by only using half training data than state-of-the-art approaches.

Features analysis showed that using only the ten most discriminative, representing all three categories, works well. Some features based on strong evidence on others tasks were not as useful as expected: the presence of the entity in the title (which is known for being a good summary of the document) and position of the entity in documents.

A lot of things remain unexplored: the time dimension needs further investigation (is the profile of the entity must be updated over time?, does burstiness help? only for some kind of documents? ...); the characteristics of each type of document might to be considered separately; lastly, is filtering highly relevant documents is helpful for automatic knowledge base population tasks like slot-filling?

6. REFERENCES

[1] R. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. *Proceedings of the 11th Conference EACL*, 2006.

[2] N. Cancedda, C. Goutte, J.-M. Renders, N. Cesa-Bianchi, A. Conconi, Y. Li, J. Shawe-Taylor, A. Vinokourov, T. Graepel, and C. Gentile. Kernel methods for document filtering. *Proceedings of The 11th TREC*, 2002.

[3] T. Cassidy, Z. Chen, J. Artiles, H. Ji, H. Deng, L. Ratinov, J. Han, D. Roth, and J. Zheng. Cuny-uiuc-sri tac-kbp2011 entity linking system description. *Proceedings of the Fourth TAC*, 2011.

[4] A. Davis, A. Veloso, A. da Silva, W. M. Jr., and A. Laender. Named entity disambiguation in streaming data. *Proceedings of the 50th meeting of the ACL*, 2012.

[5] J. Frank, M. Kleiman-Weiner, D. Roberts, F. Niu, C. Zhang, and C. Ré. Building an entity-centric stream filtering test collection for trec 2012. *Proceedings of The 21th TREC*, 2012.

[6] S. Gottipati and J. Jiang. Linking entities to a knowledge base with query expansion. *Proceedings of the Conference EMNLP*, 2011.

[7] H. Ji, R. Grishman, and H. Dang. Overview of the tac2011 knowledge base population track. *Proceedings of the Fourth TAC*, 2011.

[8] B. Kjersten and P. McNamee. The hltcoe approach to the trec 2012 kba track. *Proceedings of The 21th TREC*, 2012.

[9] J. Lee. Mining spatio-temporal information on microblogging streams using a density-based online clustering method. *Expert Systems with Applications*, 2012.

[10] P. McNamee, J. Mayfield, V. Stoyanov, D. Oard, T. Xu, W. Ke, and D. Doermann. Cross-language entity linking in maryland during a hurricane. *Proceedings of the Fourth TAC*, 2011.

[11] S. Robertson and I. Soboroff. The trec 2002 filtering track report. *Proceedings of The 11th TREC*, 2002.

[12] S. Robertson, S. Walker, H. Zaragoza, and R. Herbrich. Microsoft cambridge at trec 2002: Filtering track. *Proceedings of The 11th TREC*, 2002.

[13] R. Schapire, Y. Singer, and A. Singhal. Boosting and rocchio applied to text filtering. *Proceedings of the 21st annual international ACM SIGIR*, 1998.

[14] W. Zhang, Y. C. Sim, J. Su, and C. L. Tan. Entity linking with effective acronym expansion, instance selection and topic modeling. *Proceedings of the Twenty-Second IJCAI*, 2011.

[15] Y. Zhang and J. Callan. Maximum likelihood estimation for filtering thresholds. *Proceedings of the 24th annual international ACM SIGIR*, 2001.