Named Entity Filtering Based on Concept Association Graphs

Oskar Gross[1], Antoine Doucet[2], and Hannu Toivonen[1]

[1] Department of Computer Science
P. O. Box 68 (Gustaf Hällströmin katu 2b)
FI–00014 University of Helsinki
Finland
first.last@cs.helsinki.fi

[2] Normandy University – UNICAEN
GREYC, CNRS UMR–6072
F–14032 Caen Cedex
France
first.last@unicaen.fr

Abstract. In this paper, we introduce a novel technique for named entity filtering, focused on the analysis of word association networks. We present an approach for modelling concepts which are distinctively related to specific named entity. We evaluated our approach in the context of the TREC Knowledge Base Acceleration track, and we obtained significantly better performance than the top-ranked systems. For this task, given the set of all named entities and nouns, our approach proved better-performing for named entity filtering than the baseline SVM classifier. This performance is the result of the ability to disambiguate entities, by taking into account the concepts relevant to a specific named entity.

1 Introduction

In this paper we will demonstrate a method for detecting documents which are related to a target named entity. The complexity of the task lies in the fact, that a named entity and a document could be related even when the named entity is not explicitly mentioned in the document. Moreover, it is possible that two absolutely different entities have the same name (e.g. *Queen* might refer to a British rock band, a British women's magazine or a subway station in Toronto).

Knowledge bases (e.g. Wikipedia) collect, structure and validate information about certain entities or events. At the moment articles in the knowledge bases are managed by humans and new information is added to the article with some delay. According to Frank et al. [1] the median of the updates delay in Wikipedia is 356 days. Detecting automatically news stories, which are novel and relevant to Wikipedia articles would decrease human labour a lot. In this paper we are focusing on detecting documents, which are relevant to a news story and omit the novelty aspect. In addition to knowledge base acceleration, some examples of other potential applications are media monitoring, topic mining and advertising.

We propose a graph based method for relating documents to target named entities. The fundamental idea of the method is to model a named entity by analysing its co-occurring concepts. We will provide a methodology for creating named entity specific graphs, which we use for filtering documents. We will evaluate our methodology by using data provided by NIST during the TREC *Knowledge Base Acceleration* (KBA) track in 2012. The main motivation for this task is to detect documents about entities, which may contain information to be added to the knowledge bases.

The rest of the paper is organized as follows: in the next section we will introduce the related work. In Section 3 we will introduce the method for generating concept graphs. How to filter the documents using the proposed method will be described in Section 4. We evaluate our method and compare its results to the state of art in Section 5. Finally, conclusions are drawn in Section 6.

2 Related Work on Named Entity Filtering

Named-entity filtering, from a stream of news data, is related to several fields where discovering and following-up on events concerning a given topic is especially valuable. In all these fields, the ability to identify named-entities is an essential performance enhancer.

Followingly, this task concerns diverse fields of information retrieval, such as news surveillance [2], entity linking [3] and text categorization [4]. In this section, we will focus on the closest and most significant papers, notably on the approaches developed during the recent TREC KBA track, whose first round in 2012 [1] focused specifically on the task of named-entity filtering.

News Surveillance. The task of news surveillance is to give alerts for all the events related to a given domain of interest. For instance, health agencies (e.g., the World Health Organization) wish to be informed of every case of occurrence of a transmittable disease, as close as possible from the moment when it occurred [2]. Other typical fields of application lie in the field of intelligence, and in finance, where the era of high frequency trading turned the apprehension of news milliseconds earlier into a decisive advantage. However, most approaches are strongly domain-dependent, requiring thousands of syntactic patterns to detect relevant news alerts [5].

Entity Linking. Entity Linking is the task of automatically linking phrases occurring in a document to entries in a knowledge base. Several comparative evaluation competitions have run in the recent past, testifying on the great progress achieved (INEX's Link-the-Wiki [6], Text Analysis Conference's Knowledge Base Population (KBP) [7]). Entity linking is nowadays a well-understood problem, that paves one way leading towards named-entity filtering : once the named-entities are marked within a text, it "only" remains to compute the centrality and relevance of the named entity: is it the main topic of the document, or is it simply mentioned?

Many of the methods presented in the TREC KBA track follow up from entity linking. This is natural, since the corpus was provided with pre-extracted named-entities.

Liu and Fang [8] presented one of the best performing approaches of the KBA track, by building "entity profiles". By fetching a snapshot of the Wikipedia, and considering the anchor text of all internal Wikipedia links as related entities, they defined a wider representation of named entities.

Araujo et al. [9] underlined that 4% of the Wikipedia citations do not mention the Wikipedia they are cited by. This motivates their focus on the detection of documents that do not mention a named entity that is yet central to it. To achieve this, they fed their model with the Google Cross-Lingual Dictionary (GCLD) [10], a ready-made resource associating Wikipedia entries to strings. As the TREC KBA topics are named-entities for which a Wikipedia entry is defined, they could replace the topics with the strings returned by the GCLD. With adequate parameters, the technique obtained the best performance for centrality and relevance.

Text Categorization. Text categorization is the task of assigning categories to a text, given a training set of text-category assignments. Text filtering is the special case when there is only one category, and the classifier is only to decide whether a given text belongs to it, or not. Such a categorization is usually led based on word term features, and the best-performing technique in the state of the art is the well-known SVM [11].

Kjersten and McNamee [12] hence proposed to filter the document sets, using the SVM classifier over a set of features composed of the named entities provided by the TREC KBA organizers. Positive examples from the training set were those marked as central. All the others were considered negative. The technique proved that this was achievable, and it obtained the best and second-best performance (out of 40 runs) for centrality.

Other approaches. The approaches presented at TREC KBA 2012 can essentially be split into two categories [1]: those that exploit rich features from a Knowledge Base (Wikipedia or Google Dictionary) and those that focus on machine learning techniques (such as SVM).

Unlike the approaches from the first category, our technique is endogenous, that is, it does not make use of any resources that are not present in the corpus. Hence, it can easily generalize accross domains and languages (even though, the latter was not yet verified).

To the best of our knowledge, no recent techniques have been proposed that would rely on the construction and exploitation of concept association graphs. The closest example was introduced by Gamon [13]. He adressed the problem of novelty detection by building an association graph connecting sentences and sentence fragments, and chose to exploit a number of graph-based features that were assumed to be good indicators of novelty. The method tied with the best techniques presented in the TREC novelty track 2 years earlier [14], but the authors himself questioned the significance of the improvement.

In the following sections, and in the light of related work, we will introduce our approach in full details.

3 Named Entity Modelling

Our method is based on the idea, that a news item is related to a named entity when both of them are related to the same concepts. Thus our approach consists of two steps:

1. We calculate which concepts are related to each other, building an association graph of named entities;
2. For each news story, and for each named entity in a query (or TREC topic), we calculate the overlap between the concepts related to the named entity and those related to the news story.

In the rest of this section, we will detail, in chronological order, the different steps in which we process document stream data so as to build our background concept association graph.

3.1 Selecting Concepts

To build our concept graph, the very first step is naturally to select the concepts. We will use TREC data which contains part-of-speech tags, lemmatized forms of the words and is annotated with Stanford NER [15]. The annotation of Stanford NER identifies whether each single word is itself or is part of a named entity and tags it with a type. For the data to fit our purposes, we post-processed the resulting set of named entities as follows :

1. Concatenate each named entity parts with an underscore (adjacent words with the same type (organization, person etc) are concatenated together);
2. Remove all words which are not nouns.

We extract nouns and named entities from the documents and discard everything else. This choice is motivated by nouns and named entities being conceptually more basic than concepts referred to by verbs or prepositions [16]. Of course, we lose some information with this step. As a final step, we lower-case and leave the lemmatized form of the words.

3.2 Building the Concept Graph

We build the concept graph, based on learning data, containing annotations indicating which named entities and which news stories are truly related. The graph generation consists of two steps: (1) calculating the co-occurrence graph using the documents and (2) cleaning the graph, by removing unnecessary edges and nodes. We will next give an overview of the graph construction, which is follows the same as principles as in Gross et.al in [17].

The first step is based on log-likelihood ratio calculation. Consider the set of documents, which are connected (by annotation) to named entity n, by $d \in C_n$. We will consider a document d as bag of sentences S_d and each sentence as bag of words $T_d \in S_d$. The set of all words is $T = \bigcup T_d$.

We analyze word co-occurrences on the granularity of sentences, since words which are in one sentence generally have a stronger relation to each other [18].

The concept graph $G_n = (V_n, E_n, W_n)$ is a weighted, undirected graph with nodes V_n, edges $E_n \subset V_n \times V_n$, and edge weights $W_n : V_n \times V_n \to \mathbb{R}_+$. For notational convenience, we assume $W(e_1, e_2) = 0$ if there is no edge between e_1 and e_2.

Construction of the graph then starts by using all terms in the corpus C_n as nodes, i.e., $V = T$.

We use the log likelihood ratio (LLR) to measure the strength of an association between two terms [19]. In [17] we showed that the co-occurrences, as measured by LLR do make sense, though other word association measures would probably be equally suitable.

LLR measures how much the observed joint distribution of terms x and y differs from their distribution under the null hypothesis of independence, i.e., how strong is the association between them. Edges are constructed for term pairs $\{e_1, e_2\}$ in T that have a strong log-likelihood ratio $LLR(e_1, e_2)$.

In other words, we in principle compute LLR for the union P of all the pairs of terms in all sentences of the corpus:

$$P = \bigcup_{d \in C_n} \bigcup_{s_d \in d} s_d \times s_d. \tag{1}$$

3.3 Cleaning up the Graph

The goal of the graph cleaning process is to remove edges and nodes which, at this point, we find unnecessary. We are interested in leaving only such associations, which are directly related to named entities. For this, we define N as the combinations of the different parts of each named entity. Consider a named entity "Annie_Laurie_Gaylor". For this named entity the possible combinations are $N = \{$"Annie_Laurie_Gaylor", "Annie_Laurie", "Annie_Gaylor", "Laurie_Gaylor", "Annie", "Laurie", "Gaylor"$\}$. In the next step we leave only nodes which are associated to the parts of the named entity, i.e. $e_1 \in N \vee e_2 \in N$.

Our experience showed, that there are nouns, which appear in all the named entity graphs. For reducing some amount of the noise, we remove all such nodes, which appear in all the different named entity graphs. Let Γ denote the set of all named entity graphs. Then we will construct a set of nodes which are found in all graphs as

$$U = \bigcap_{G_n \in \Gamma} V_n(G_n).$$

These nodes will be removed from all the graphs:

$$G_n = (V \setminus U, \{e \in E : e_1 \notin U \wedge e_2 \notin U\}, W).$$

In the next section we will show, how we utilise these graphs for detecting the news stories which are related to a given topic.

4 Document Filtering with Concept Graphs

4.1 Principles

To be able to rank documents with respect to the NEs of any given TREC KBA topic, consisting of one or more named entities, it remains to design a way to compute relevance scores based on our graph model.

We do so by relying on the concept of word co-occurrence, with the following principles in mind, on what we expect a more interesting document to be like. First it is reasonable to assume, that the concepts in the document should intersect with the concepts which are strongly related to the named entity. On the other hand, as the named entity could appear in many different contexts (e.g. the president of the USA could be related to financial, political, arts & entertainment topics et cetera), we should not penalize a document for *not* being related to some neighbours.

4.2 Document Relevance Evaluation

To calculate the relevance score of a given document, w.r.t. a given TREC KBA topic, we proceed as follows. We post-process the document exactly as we described in Section 3.1 and calculate the named entity specific graph models. Then we use the weights in the named entity graphs to calculate the relevance of a document.

As the first part is covered in the previous section, we will hereby describe the second step.

Documents relevance is calculated by measuring how strongly words in the document are connected to the named entity of interest. Let us consider the target document d_t which contains the words w_t.

For a named entity n we calculate the relevance status value RSV for an incoming document, given the entity graph G_n as:

$$RSV(G_n, d_t) = \frac{1}{|w_t|} \sum_{w \in w_t} \sum_{v \in V_n} W(w, v)$$

which is the average edge weight of the words in the named entity graph. The rationale for calculating RSV as the average of the edge weights is to reduce the impact of outliers. Indeed, we believe that averaging over all the edges represents better the *general* match between the document and the named entity, than, e.g., summing up or taking the maximum weight. This is the case when the overlap between the document and the named entity graph is small (e.g. 3 or 4 nodes) and one node is strongly connected to the entity graph and other nodes are weakly connected to the named entity.

5 Evaluation

In this section we will describe the evaluation methodology and present the subsequent results.

5.1 Method

The KBA 2012 evaluation data consists of 57,750 human-generated judgements rating the relevance of documents to target entities. The KBA stream corpus of 462M documents covers 4,973 contiguous hours of news, blogs, and forum posts. It includes dozens of languages beside English. In our task we use a subset of the data – documents with a reasonable chance to be written in English that have been automatically POS tagged. This subset contains roughly 367M documents.

The data spans over 8 months - from October 2011 till May 2012. The data is divided into two sets by using a *cut-off* date, which is January the 1st. Documents published before the cut-off date are used as the training set and documents after the cut-off date form the test set.

Each article in the annotation set is scored in two categories - *relevant* and *central*. For each entity, the document *central* score is high if the respective entity is the central topic of the document. The *relevant* score is high if the document is indirectly relevant to a certain named entity. In the evaluation no pooling of the TREC participants results is used.

The accuracy of the method is calculated on the test set by using the articles which are annotated. However, in the scoring phase, the algorithm does not know which articles are annotated and not, and it therefore needs to go through all 367M documents.

The methods are evaluated by using the standard information retrieval measures – precision, recall and their harmonic mean F_1.

5.2 Graph Based Model

The graph models for each named entity are created by using the methodology described in Section 3. The similarity score for each document and named entity pair is calculated by using the method given in Section 4.

Baseline

For comparing our methodology, we will use a standard machine learning algorithm, support vector machines (SVM) as a baseline. SVM has been shown to be successful in text categorization [20] and document filtering is a special case of text categorization with two categories.

For each named entity we calculate the entity specific SVM model by using the annotation data. We use the same feature-set as we use in the graph based models. In total there are 13,111 features. The features are used as binary features, representing whether a word is found in the document or not.

We carried out our SVM experiments with SVM light toolkit [11]. Following good data analysis practices, we divided the training set into two parts – 80% of the documents as a training set and 20% of the news into test set. We used these two sets to estimate the performance of our method and also to analyze SVM scores for different documents. We observed that the meaningful values for SVM classification were between -2 and 2, thus we scaled this range to be between 0 and 1000.

The SVM models used for scoring after the cut-off date were trained on the whole dataset which was published before the *cut-off* date (i.e. 1st of January 2012).

Results

For measuring the performance of our method, we will compare it to the best scoring methods at the TREC 2012 KBA track – `CWI-google_dic_31` proposed by Araujo et al. [9] and `hltcoe-wordNER` proposed by Kjersten and McNamee [12]. `CWI-google_dic_31` uses Google Cross Lingual dictionary and the `hltcoe-wordNER` uses SVM for document filtering, very similarly to our baseline.

For measuring, whether the accuracy of the different methods is significant, we approximated the pairwise p-values between performance values of the methods. The statistical significance is calculated by using two samples – for each method we chose the cutoff value for which the average F_1 measure was maximized. The sample consists of all the topics F_1 measures for respective cutoff. We then use the Wilcoxon Rank Sum test for testing whether the difference between samples is significant.

The results are summarized in Table 1. We can observe that we are doing better than the baseline methodology in both categories. In the *central* category our method is a bit worse than the best performing method of the KBA track in TREC (`hltcoe-wordNER`), although not significantly. On the other hand, our methodology is the best perfomer in the *central+relevant* category by beating `CWI-google_dic_31`. The reason for such good accuracy in this category can be related to the approach of our methodology – we put strong emphasis on the context, where the named entity appears.

While our graph-based approach outperforms all the other methods in the *central+relevant* category, it is important to emphasize that a strength of our method is that it is resource free and relies only on the corpus. It should hence be much easier to generalize to other data sets, e.g., written in other languages.

Future Work

One of the current shortcomings of our graph based models is that they are static. Intuitively it is reasonable to assume that the accuracy of the graph models decreases over time, as the topics about a certain entity may also change over time. As an example one could consider a sportsman, who at different times, can be related to documents about Olympics, gold medals and doping.

Table 1. The F_1 measures for central and central+relevant. The asterisk marks the F_1-measures which are significantly ($p < 0.05$) different (in the same category) from the graph based method.

Method	**Central**	**Central+Relevant**
Baseline (SVM)	0.327	0.569*
CWI-google_dic_31	0.291	0.637
hltcoe-wordNER	**0.359**	0.494*
Named Entity Graph	0.341	**0.691**

To test the hypothesis, we calculated the precision of our method over time by using the test set annotation files. The result can be observed in Figure 5.2.

As shown in the figure, the precision is high in the beginning, but it gets lower over time, suggesting that our model becomes more and more outdated. Observe, that the drop from 0.4 to ≈ 0.3 is rapid, approximately two weeks, and after this the precision decreases in a more constant rate. This observation is a good hint for pointing our future work towards updating the graph models while analysing the stream.

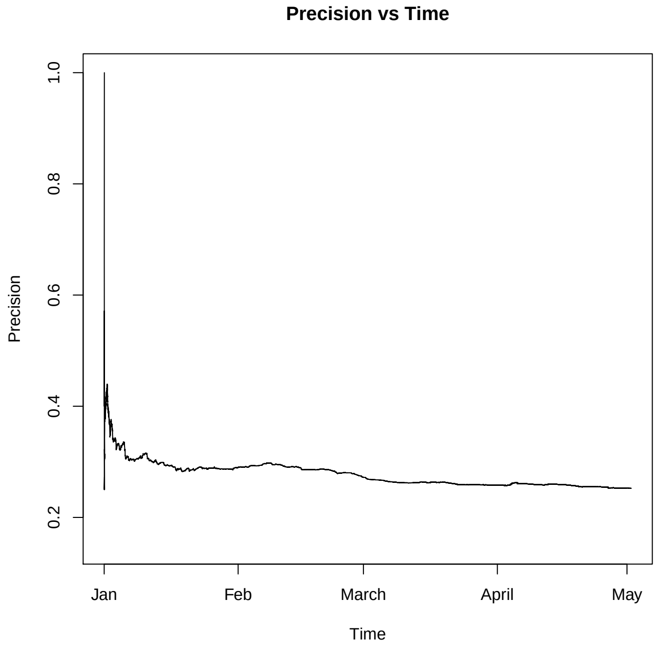

Fig. 1. Decrease of the precision of the graph based method over time.

6 Conclusion

In this paper we have demonstrated a corpus based approach for modelling named entities. When designing the method, we have considered the following aspects. First, we find it important to be independent from language as much as possible and we use fairly simple methods in order to get rid of some of the obvious noise. Secondly, we have chosen a graph based approach due to the interpretable and easily expandable nature of the models.

We have implemented and experimented with our approach with encouraging results. Immediate future work will focus of implementing the update of the background graph, which is currently static, based only on the training data. In context of stream data, using static models is inadequate. The main aspect of the named entities is, that they evolve in time. We believe that taking into account the temporal aspect of the named entities could give a lot of improvement of the performance of the system.

Acknowledgements: This work has been supported by the Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland.

References

1. Frank, J.R., Kleiman-Weiner, M., Roberts, D.A., Niu, F., Zhang, C., R, C., Soboroff, I.: Building an Entity-Centric Stream Filtering Test Collection for TREC 2012. In Voorhees, E.M., Buckland, L.P., eds.: Proceedings of the 21st Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, November 6-9, 2012, National Institute of Standards and Technology (NIST) (2012)
2. Linge, J., Steinberger, R., Weber, T., Yangarber, R., van der Goot, E., Al Khudhairy, D., Stilianakis, N.: Internet surveillance systems for early alerting of threats. Eurosurveillance **14** (2009)
3. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In Silva, M.J., Laender, A.H.F., Baeza-Yates, R.A., McGuinness, D.L., Olstad, B., Olsen, y.H., Falco, A.O., eds.: CIKM, ACM (2007) 233–242
4. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34** (2002) 1–47
5. Steinberger, R.: A survey of methods to ease the development of highly multilingual text mining applications. Language Resources and Evaluation (2011) 1–22
6. Huang, D.W., Xu, Y., Trotman, A., Geva, S.: Overview of INEX 2007 link the wiki track. In Fuhr, N., Kamps, J., Lalmas, M., Trotman, A., eds.: Focused Access to XML Documents. Springer-Verlag, Berlin, Heidelberg (2008) 373–387
7. Ji, H., Grishman, R.: Knowledge base population: successful approaches and challenges. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 1148–1158
8. Liu, X., Fang, H.: Entity Profile based Approach in Automatic Knowledge Finding. In Voorhees, E.M., Buckland, L.P., eds.: Proceedings of the 21st Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, November 6-9, 2012, National Institute of Standards and Technology (NIST) (2012)

9. Araujo, S., Gebremeskel, G., He, J., Bosscarino, C., de Vries, A.: CWI at TREC 2012, KBA Track and Session Track. In Voorhees, E.M., Buckland, L.P., eds.: Proceedings of the 21st Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, November 6-9, 2012, National Institute of Standards and Technology (NIST) (2012)
10. Spitkovsky, V.I., Chang, A.X.: A cross-lingual dictionary for english wikipedia concepts. In Chair), N.C.C., Choukri, K., Declerck, T., Doan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., eds.: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, European Language Resources Association (ELRA) (2012)
11. Joachims, T.: Making large-scale support vector machine learning practical. In Schölkopf, B., Burges, C.J.C., Smola, A.J., eds.: Advances in kernel methods. MIT Press, Cambridge, MA, USA (1999) 169–184
12. Kjersten, B., McNamee, P.: The HLTCOE Approach to the TREC 2012 KBA Track. In Voorhees, E.M., Buckland, L.P., eds.: Proceedings of the 21st Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, November 6-9, 2012, National Institute of Standards and Technology (NIST) (2012)
13. Gamon, M.: Graph-based text representation for novelty detection. In: Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing, New York City, Association for Computational Linguistics (2006) 17–24
14. Soboroff, I.: Overview of the trec 2004 novelty track. In Voorhees, E.M., Buckland, L.P., eds.: TREC, National Institute of Standards and Technology (NIST) (2004)
15. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. ACL '05, Stroudsburg, PA, USA, Association for Computational Linguistics (2005) 363–370
16. Gentner, D.: Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. BBN report; no. 4854 (1982)
17. Gross, O., Toivonen, H., Toivanen, J.M., Valitutti, A.: Lexical creativity from word associations. In: Knowledge, Information and Creativity Support Systems (KICSS), 2012 Seventh International Conference on. (2012) 35 –42
18. Miller, G.: Wordnet: a lexical database for english. Communications of the ACM **38** (1995) 39–41
19. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational linguistics **19** (1993) 61–74
20. Joachims, T.: Text categorization with suport vector machines: Learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning. ECML '98, London, UK, UK, Springer-Verlag (1998) 137–142