Term Association Analysis for Named Entity Filtering

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. This paper describes the participation of the Universities of Helsinki and Caen in the first round of the TREC Knowledge Base Acceleration track[3]. The task focused on filtering a stream of documents relevant to a set of entities. Our approach uses word co-occurrence graphs for modelling the named entities. We submitted two runs that achieved an average F-measure superior to the mean of all submitted runs. The best of those runs ranked in the top 5 runs for both the central and relevant F-measures, out of a total of 43 runs submitted by 11 institutions. As our runs were the produce of a first implementation of our approach, these preliminary results are very supportive of our idea to use concept graphs for modelling named entity relations.

1 Introduction

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 articles with some delay. According to Frank et al. [1] the median delay of the updates in Wikipedia is 356 days. Automatically detecting news stories, which are novel and relevant to Wikipedia articles would considerably decrease the amount of human labour needed to perform this task. In addition to knowledge base acceleration, other potential applications include 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 provide a methodology for creating named entity specific graphs, which we use for filtering documents.

[3] http://trec-kba.org/

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

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.

3 Named Entity Graphs

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

1. Calculate which concepts are related to each topic item (named entity graphs);

2. Calculate, for each news story, the overlap between the concepts related to the named entity and those related to the news story.

In this report we present a very simple approach: due to time constraints, our runs were only meant to eavaluate the potential of our approach. The generalization of the method to graphs, is here to provide a foundation on which we can base our future work. Indeed, our TREC KBA 2012 experiments used only the 1st level associations of the named entity graphs.

Stanford NER processing. Before using the news stories, we processed the Stanford NER data as follows:

1. Concatenate named entity names with underscore (if the type (organization, person etc) of the previous word is the same as the current word, they are concatenated together)
2. Remove all words which are not nouns

We extract nouns and named entities from the documents and discard everything else. In addition to simplicity, this choice is motivated by nouns and named entities being conceptually more basic than concepts referred to by verbs or prepositions [15]. We then lowercase and lemmatize all the words.

Named entity graphs. The named entity graphs are calculated by using the annotation data and all the news stories before the cutoff period. The graph generation consists of two steps: (1) we calculate the co-occurrence graph using the documents; (2) we clean the graph, by removing unnecessary edges and nodes.

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 have a strong relation to each other [16].

The named entity 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 $W_n : V_n \times V_n \to R_+$. For notational convenience, we assume $W(e_1, e_2) = 0$ if there is no edge between e_1 and e_2.

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

We use log-likelihood ratio (LLR) to measure the strength (or unexpectedness) of an association between two terms [17].

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 surprising is their association. Edges are constructed for each term pair $\{e_1, e_2\}$ in T that has a high-enough LLR value.

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

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

Cleaning the graphs. The goal of the graph cleaning process is to remove unncessary edges and nodes. We are interested in keeping only associations that are directly related to the named entity. In this aim, we first calculate the combinations N of the different parts of the 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 such edges, for which $e_1 \in N$ or $e_2 \in N$.

Our experiments showed, that there are nouns, which appear in all the named entity graphs. For overcoming this noisiness problem, we removed all nodes that appeared in every single named entity graph, since their discriminative power is subsequently very weak. Let the set of all topic graphs be $G_n \in \Gamma$. We construct the set of nodes which are found in all the graphs as:

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

The following nodes are hence removed from all the graphs:

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

For demonstration purposes, we extracted the highest weighted edges and respectively adjacent nodes from two graphs – one is the graph of the pianist Boris Berezovsky, and another is for the businessman Boris Berezovsky. The resulting graphs can be seen on Figure 1 for the pianist and Figure 2 for the businessman.

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

4.1 Principles

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

We did 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 :

- the concepts in document should intersect with concepts related to NE;
- the other (unrelated) neighbours should not be penalized for.

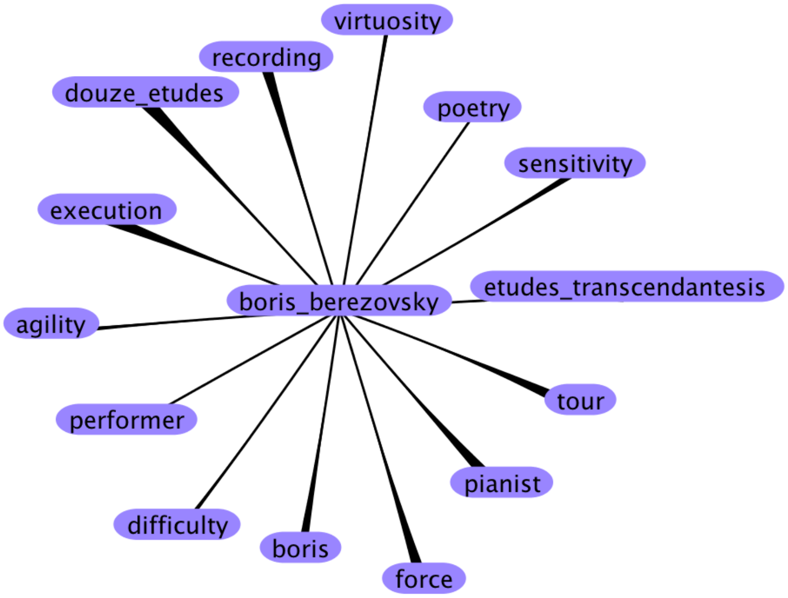

Fig. 1. The graph for the pianist Boris Berezovsky. We have omitted the edge weights for the sake of clarity of the figure.

4.2 Document Relevance Evaluation

Following the preprocessing step described in the previous section, we use the weights in the named entity graph to calculate the relevance of documents as follows.

Our main idea is that document relevance is calculated by measuring how strongly words in the document are connected to the named entity. Let us consider the target document d_t, containing the words w_t.

For a given named entity n we calculate the relevance status value r between the document d_t and 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.

To produce our 2 runs,we used two different thresholding methods:

1. For the first run, *helsinki-disgraph50*, we used a fixed, rule-of-thumb threshold of 30;
2. For the second run, *helsinki-disgraph250*, we used the mean value of the weights of the entity graph of the current entity.

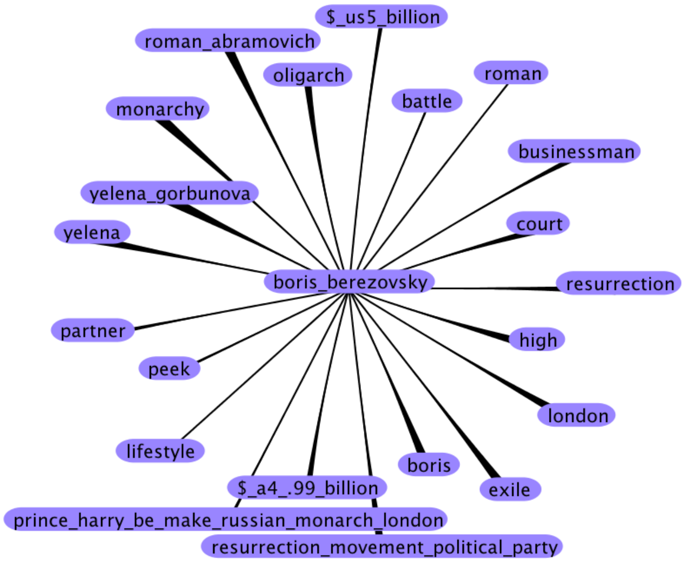

Fig. 2. The graph for the businessman Boris Berezovsky. We have omitted the edge weights for the sake of clarity of the figure.

5 Run Description

5.1 Testing and Results

To get a first glimpse of the suitability of our methodology, we experimented with the training data provided, using the following procedure.

We first split the training data into a learning set and a test set, containing, respectively, 80% and 20% of the training documents. Splitting the data was done according to chronology: the news stories assigned to the training set occurred before those assigned to the test set.

To estimate the accuracy of our method, we first built the graphs for each target topic. Then, for each document d, we ran the following procedure through the test data:

1. Compute the relevance score of document d (as described in Section 4.2)
   (a) If the document was assessed by the annotators, store the annotator decision and the score;
   (b) If $score_d > 0$, store the value;
   (c) Otherwise, ignore the document.

Followingly, we obtained a document sample containing all annotated documents, and all documents with a score higher than 0, that is, all those documents containing words that are directly related to the corresponding TREC KBA topic.

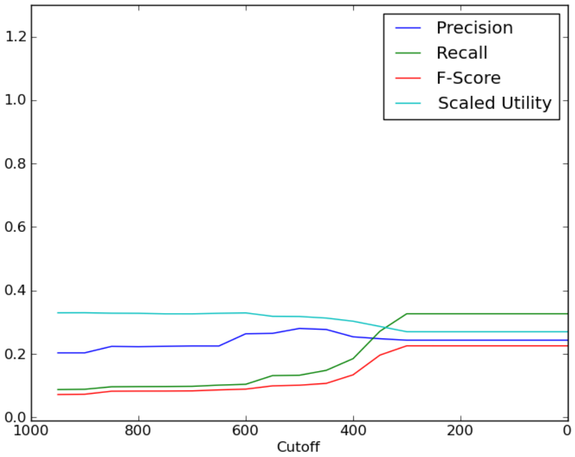

Fig. 3. Run "disgraph": Precision, Recall, F-measure and Scaled Utility w.r.t. cut-off value

We then split this dataset into two disjoint subsets named *related* and *unrelated*. The *related* set contained the documents for which the annotation decision was set to "are related", while all the other documents were thrown into the *unrelated* set. The corresponding prediction was encouraging, with an area under the ROC curve around 0.77.

5.2 Submitted Runs and Results

Using this methodology, we submitted two runs to the KBA track, *disgraph* and *disgraph2*. The main methodology for both of the methods was the same, and, as mentioned earlier, the only difference was in the threshold selection method. For *disgraph* the threshold was manually set to 30 and for *disgraph2* the threshold was equal to the mean value of each named entity graph.

We calculated the named entity graphs on the whole training set, i.e., using all the data before the cutoff date, which was 31.12.2011 23:59:59. The named entity graphs were then used for scoring documents by using the two different thresholding methods.

The performance of our two runs, *disgraph* and *disgraph2*, are summarized in Table 1. The detailed performance of both runs over different cut-off values is given respectively in Figures 3 and 4.

The mean average F-measure for the KBA-track was 0.2066 and the best run's average F-measure was 0.4263. Taking into account documents in which topic named entities were marked both relevant and central, the run *disgraph2* was ranked 3rd (out of 43) in terms of macro-averaged F-score, and 5th in terms

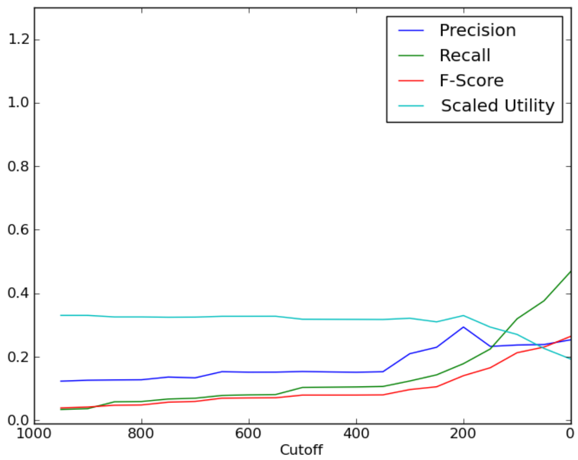

Fig. 4. Run "disgraph2": Precision, Recall, F-measure and Scaled Utility w.r.t. cut-off value

Table 1. Evaluation of our two official runs

Run ID	Scaled Utility	Precision	Recall	F-measure
disgraph	0.2699	0.24323	0.3267	**0.2256**
disgraph2	0.1929	0.2535	0.4688	**0.2645**

of micro-average F-score (see the overview paper of the KBA track for further details [1]).

6 Conclusion

In this paper we proposed a method for filtering documents according to named entities. The method relies on finding the concepts, i.e., nouns and other named entities, which are related to the respective named entity.

The method was designed with the goal to be as language independent as possible. Another aspect we took into account was the interpretability and the possibility to generalize the method in the future.

The method works well, performing amongst the top 5 runs out of 43, in terms of F-measure over documents in which the topic was judged relevant and central. This is a very encouraging result for further exploring the idea of modeling named entity relations through concept association graphs.

Essential future work is to become able to update the named as the stream is analyzed. In our current implementation, the named entity graph is statically based on documents prior to the cutoff date of 31.12.2011 at 23:59:59, and we observed a clear and steady decrease in precision as the documents processed were getting farther from this date.

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. [18]
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, .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.: 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. [18]
9. Araujo, S., Gebremeskel, G., He, J., Bosscarino, C., de Vries, A.: CWI at TREC 2012, KBA Track and Session Track. [18]
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.: 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. [18]
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. Gentner, D.: Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. BBN report; no. 4854 (1982)
16. Miller, G.: Wordnet: a lexical database for english. Communications of the ACM **38** (1995) 39–41

17. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational linguistics **19** (1993) 61–74
18. 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)