fokas: Formerly Known As
A Search Engine Incorporating Named Entity Evolution*

Helge HOLZMANN *Gerhard GOSSEN* *Nina TAHMASEBI*
L3S Research Center, Appelstr. 9a, 30167 Hannover, Germany
{ holzmann, gossen, tahmasebi }@L3S.de

ABSTRACT
High impact events, political changes and new technologies are reflected in our language and lead to constant evolution of terms, expressions and names. This makes search using standard search engines harder, as users need to know all different names used over time to formulate an appropriate query. The *fokas* search engine demonstrates the impact of enriching search results with results for all temporal variants of the query. It uses NEER, a method for named entity evolution recognition. For each query term, NEER detects temporal variants and presents these to the user. A chart with term frequencies helps users choose among the proposed names to extend the query. This extended query captures relevant documents using temporal variants of the original query and improves overall quality. We use the New York Times corpus which, with its 20 year timespan and many name changes, constitutes a good collection to demonstrate NEER and *fokas*.

TITLE AND ABSTRACT IN GERMAN

fokas: Früher bekannt als
Eine Suchmaschine mit Einbindung von Namensevolution

Wichtige Ereignisse, politische Veränderungen und neue Technologien spiegeln sich in unserer Sprache wieder und führen zu einer ständigen Evolution von Begriffen, Ausdrücken und Namen. Dies erschwert die Suche mit herkömmlichen Suchmaschinen, da Nutzer zur Formulierung einer Anfrage sämtliche Namen kennen müssen, die im Laufe der Zeit verwendet wurden. Die Suchmaschine *fokas* zeigt den Einfluss des Anreicherns der Suchergebnisse mit den Ergebnissen für allen zeitlichen Varianten des Suchbegriffs. Sie verwendet NEER, eine Methode zur Erkennung von Namensevolution. NEER erkennt für jeden Suchbegriff alle zeitlichen Varianten und präsentiert diese dem Nutzer. Ein Termfrequenz-Diagramm ergänzt die Ergebnisse, um Nutzern bei der Wahl zwischen den vorgeschlagenen Namen zur Erweiterung der Anfrage zu unterstützen. Diese erweiterte Anfrage findet relevante Dokumente, die nur eine zeitliche Variante der ursprünglichen Anfrage verwenden, und verbessert dadurch die Gesamtqualität. Wir verwenden den Korpus der New York Times, der mit seiner Zeitspanne von 20 Jahren und vielen Namensänderungen eine gute Kollektion zur Demonstration von NEER und *fokas* ist.

KEYWORDS: Named Entity Revolution Recognition, unsupervised, NEER, search engine.

GERMAN KEYWORDS: Namensevolutionserkennung, nichtüberwacht, NEER, Suchmaschine.

*This work is partly funded by the European Commission under ARCOMEM (ICT 270239)

Proceedings of COLING 2012: Demonstration Papers, pages 215–222,
COLING 2012, Mumbai, December 2012.

1 Introduction

Do you remember the bright yellow Walkman, Joseph Ratzinger or Andersen Consulting? Chances are you do not, because as the world around us changes, new terms are created and old ones are forgotten. High impact events, political changes and new technologies are reflected in our language and lead to constant evolution of terms, expressions and names. Most everyday tasks, like web search, have so far relied on the good memory of users or been restricted only to the current names of entities. As the web and its content grow older than some of its users, new challenges arise for natural language tasks like information retrieval to automatically determine relevant information, even when it is expressed using forgotten terms.

Language evolution is reflected in documents available on the web or in document archives but is not sufficiently considered by current applications. Therefore, if the users do not know different names referencing the same entity, information retrieval effectiveness becomes severely compromised. In this demonstration we present *fokas*, a search engine that knows about names used at different points in time to refer to the same named entity (called *temporal co-references*) and uses them to help users find all relevant documents. *fokas* is based on the NEER method (Tahmasebi et al., 2012b). NEER is an unsupervised method for named entity evolution recognition independent of external knowledge sources. It finds time periods with high likelihood of named entity evolution. By analyzing only these time periods using a sliding window co-occurrence method it captures evolving terms in the same context and thus avoids comparing terms from widely different periods in time. This method overcomes limitations of existing methods for named entity evolution and has a high recall of 90% on the New York Times corpus. Furthermore, using machine learning with minimal supervision leads to a precision to 94%.

In this demonstration we use the outcome of NEER to improve search on the New York Times corpus (NYTimes) by identifying temporal co-references and suggesting these to the user as possible query expansions. NYTimes serves as a good corpus because of its long time span (1986–2007), the wide range of topics and the high quality of the texts.

2 The NEER Method

Identifying Change Periods We use the Kleinberg algorithm (Kleinberg, 2003) to find bursts related to an entity. We retrieve all documents in the corpus containing the query term, group them into monthly bins and run the burst detection on the relative frequency of the documents in each bin. Each resulting burst corresponds to a significant event involving the entity. However, these bursts do not necessarily correspond to a name change. By choosing the `topB` strongest bursts we expect to find a subset of bursts which also captures change periods. We denote each **change period** p_i for $i = 1, \ldots, \texttt{topB}$.

Creating Contexts After identifying change periods p_i for an entity w we use all documents $\mathbf{D}_{w_i}$ that mention the entity or any part of it and are published in the year corresponding to p_i. We extract nouns, noun phrases and named entities. All extracted terms are added to a **dictionary** and used for creating a co-occurrence graph (see Figure 1a). The co-occurrence graph is an undirected weighted graph which links two dictionary terms if and only if they are present in $\mathbf{D}_{w_i}$ within k terms of each other. The weight of each link is the frequency with which the two terms co-occur in $\mathbf{D}_{w_i}$. The **context** of an entity w are all terms co-occurring with w.

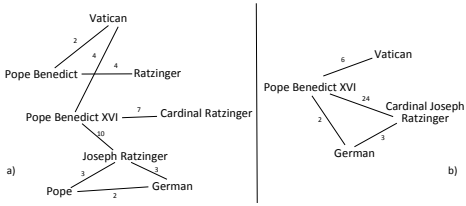

FIGURE 1: a) Example graph after creating contexts. b) After merging all direct co-references.

Finding Temporal Co-references Classes To find direct co-references (i.e., with lexical overlap) we make use of the contexts and the dictionaries. We consolidate the extracted terms by recognizing all variants of each term. The procedure for consolidation terms and groups of co-references is described in detail in (Tahmasebi et al., 2012b). During consolidation terms like *Pope*, *Pope Benedict* and *Pope Benedict XVI*, as well as *Ratzinger*, *Joseph Ratzinger* and *Cardinal Joseph Ratzinger* are considered **direct co-references** and merged. The result is displayed in Figure 1b.

Indirect Co-references Indirect co-references (i.e. without lexical overlap) are found implicitly by means of the direct co-references. After consolidation, all terms in the context are considered candidate indirect co-references. These are a mix between true indirect co-references, highly related co-occurrence phrases as well as noise. The quality of the indirect co-references is dependent on the named entity extraction, co-occurrence graph creation and filtering of the co-occurrence graph. In Figure 1b the terms *Vatican*, *German* and *Cardinal Joseph Ratzinger* are candidate co-references for *Pope Benedict XVI*. If NEER does not find any co-references for a term, all direct co-occurrences from the co-occurrence graphs (derived from the union of the change periods) are returned instead.

Filtering Temporal Co-references To remove noise and identify the true direct and indirect co-references we need to measure the temporal relatedness of terms. Unlike previous works that take temporal features into account it is not sufficient to consider relatedness over the entire time span of a collection. In Radinsky et al. (2011) the frequency of terms over times is used to capture the relatedness of terms like *war* and *peace* or *stock* and *oil*. These terms are considered related because they have similar frequencies over time. To fully capture temporal co-references we need global relatedness measures as well as a relatedness measure that captures how related terms are during the time periods where they can be related at all. To this end we allow a relatedness measure to consider only periods where both terms occur. In all cases we use the normalized frequencies. Details can be found in (Tahmasebi et al., 2012b).

We consider four relatedness measures: (1) Pearson's Correlation ($corr$) (Weisstein, 2012a), (2) Covariance (cov) (Weisstein, 2012b), (3) Rank correlation (rc) and (4) Normalized rank correlation (nrc).

The two first measures are standard relatedness measures where $corr$ measures linear dependence between random variables while cov measures correlation between two random variables. The two last measures are rank correlation measures and inspired by the Kendall's tau coefficient that considers the number of pairwise disagreements between two lists. Our rank correlation coefficient counts an agreement between the frequencies of two terms for each time

FIGURE 2: *fokas* search page.

period where both terms experience an increase or decrease in frequency without taking into consideration the absolute values. The rank correlation is normalized by the total number of time periods. The normalized rank correlation considers the same agreements but is normalized with the total number of time periods where both terms have a non-zero term frequency.

Our filtering is based on machine learning. We use a random forest classifier (Breiman, 2001) consisting of a combination of decision trees where features are randomly extracted to build each decision tree. In total ten trees with three features each are constructed. We choose features from the similarity measures presented above. This means that for each term–co-reference pair (w , w_c) found by NEER we calculate the $corr$, cov, rc and nrc measures. We also use the average of all four measures as a fifth feature. Finally we classify the pair as either 1 for w_c being a correct co-reference of w or 0 otherwise to train the classifier. We use the test set presented in (Tahmasebi et al., 2012a) for training.

3 Implementation

3.1 General Functionality

fokas stands for *formerly known as* which refers to the fact that named entities change their names over time. In *fokas*, the changed names (*temporal co-references*) will be used to expand the query posed to the system. The homepage of *fokas* mimics that of a common search engine (see Figure 2) and a typical workflow consists of the following steps:

1. enter a query string into the search field.
2. start searching by clicking the search button or selecting one of the suggested terms.
3. analyze search results, including found direct and indirect temporal co-references of the query term as well as a frequency chart of the different names.
4. select one or more relevant co-references (if any are available).
5. analyze the extended search results. Results containing only the added co-references are marked with an icon. The frequencies of the new co-references are added to the frequency chart.

While entering a query string in the search field the user will get a list of suggested terms (see Figure 3). The suggested terms are the names starting with the entered query string as well as

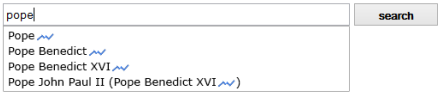

FIGURE 3: Additional query terms are suggested while typing a query.

the direct co-references found by applying NEER on the collection[1]. As shown in Figure 3, for the term *Benedict* the best matching co-references are presented (*Pope Benedict*, *Pope Benedict XVI*, *Pope John Paul II*). The first two are direct co-references for the term *Benedict*, while the last one is highly related as it refers to the previous pope. Clicking on one of the suggestions or the search button will start the search.

Next to each co-references in the suggestion box is a small graph symbol. Clicking on it will open a sidebar containing a term frequency chart and co-reference lists with all direct and indirect co-references found by NEER (see Figure 4). By clicking on one co-reference, the frequency of that term is added to the chart to help users decide the appropriateness and correctness of the chosen term.

The results of the search are presented as shown in Figure 4. The search result page has two columns. On the right hand side is the sidebar described above. The left hand side shows the search results from the articles of the New York Times corpus. These results are presented in a format similar to standard search engines, with a headline, a link to the full article and a short excerpt containing the query terms from the text of the article. The query terms (the entered query string or the selected co-references) are highlighted within the excerpt. Additionally, each result contains the publishing date of the appropriate article, which is specially relevant in the context of *fokas* and named entity evolution analysis.

fokas gives the user the ability to improve the search results by extending the query with co-references of the original query term. This can be done by selecting one or more co-references from the lists of direct or indirect co-references in the sidebar. This will immediately show the extended search results. Here all search results found through the selected co-references are marked with an icon. This lets the user directly conceive the advantage of a search augmented with the co-references of the term based on NEER, instead of only searching for the query term. The interface gives the user full control over the terms added to the query, supported by the displayed frequencies of each term.

3.2 Frequency Analysis Over Time

In addition to the search results, which are the main part of *fokas*, the frequency chart of the query is shown in the sidebar (see Figure 4, right hand side). This chart contains a graph for the query term as well as for each of the selected co-references showing the number of documents containing each term. This supports the user in selecting the co-references relevant to their query. The chart also helps to understand how NEER inferred the co-references for the query term and how the names of the appropriate entities changed over time. The blue graph shows the frequency of the query term. Each selected direct co-reference is illustrated by a green graph while the graphs for the indirect co-references are drawn in red.

[1]For efficiency reasons we currently compute co-references for a predefined set of names offline and only present these names.

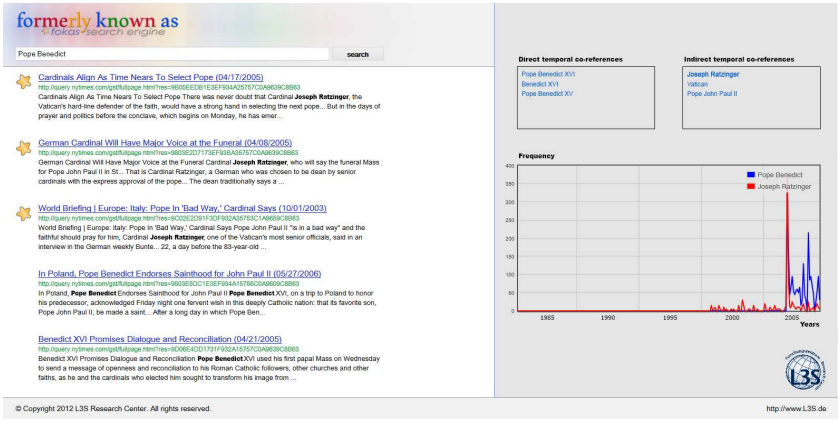

FIGURE 4: Search results enriched with results for co-references of the original query term.

As an example, the chart in Figure 4 shows the graph of the query for *Pope Benedict* after selecting the indirect co-reference *Joseph Ratzinger* to augment the results. This co-reference is a very valuable enrichment of the original query as the name *Pope Benedict* was not used before year 2005 at all. Thus, the users of *fokas* who are interested in all articles about Pope Benedict would not be able to find articles from the time before Joseph Ratzinger became Pope. Users of *fokas* will be alerted to this fact immediately and are able to take action. For example, by adding the term *Joseph Ratzinger* to the query *Pope Benedict* the user finds 34 documents (published in 2005 in the New York Times) containing only the term *Joseph Ratzinger* that would not have been found using plain keyword search.

4 Conclusion

fokas demonstrates a search engine that takes named entity evolution into account and allows users to query a document collection using temporal co-references. NEER allows us to find co-references for a wide range of terms in the New York Times corpus used by *fokas* for demonstration purposes. The lists of direct and indirect co-references and the frequency chart shown next to the search give users efficient tools for enriching their queries with their temporal variations. The highlighted search results give users a direct feedback about the improved results gained by including co-references. While *fokas* provides a well selected and filtered set of co-references based on NEER, it is not able to select the best queries for augmenting the search results automatically. *fokas* still requires interaction by the users but provides deeper insight and control, as well as transparency on how *fokas* and NEER work.

References

Breiman, L. (2001). Random forests. In *Machine Learning*, pages 5–32.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. (2011). A word at a time: computing word relatedness using temporal semantic analysis. In *WWW*, pages 337–346.

Tahmasebi, N., Gossen, G., Kanhabua, N., Holzmann, H., and Risse, T. (2012a). Named entitiy evolution dataset. Available online at `http://l3s.de/neer-dataset/`.

Tahmasebi, N., Gossen, G., Kanhabua, N., Holzmann, H., and Risse, T. (2012b). NEER: An unsupervised method for named entity evolution recognition. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, Mumbai, India.

Weisstein, E. W. (2012a). Correlation coefficient. Retrieved 2012-11-05, from `http://mathworld.wolfram.com/CorrelationCoefficient.html`.

Weisstein, E. W. (2012b). Covariance. Retrieved 2012-11-05, from `http://mathworld.wolfram.com/Covariance.html`.