SaHaRa: Discovering Entity-Topic Associations in Online News

Krisztian Balog[1] Maarten de Rijke[1] Raymond Franz[2] Hendrike Peetz[1]
Bart Brinkman[1] Ivan Johgi[1] Max Hirschel[1]
[1]ISLA, University of Amsterdam, Science Park 107, Amsterdam
k.balog,m.derijke,m.h.peetz@uva.nl, bart.brinkman,shannon.joghi,max.hirschel@student.uva.nl
[2]TrendLight Netherlands B.V., A.J. Ernststraat 595-H, Amsterdam
r.franz@trendlight.nl

ABSTRACT

We present SaHaRa, a system that helps to discover and analyze the relationship between entities and topics in large collections of news articles. We augment entity related search by including semantically related linked open data.

1. INTRODUCTION

Increasingly, news is consumed online, in ever growing volumes. With this growth has come an increasing need for intelligent access and data enrichment solutions on the publisher's side. Many efforts to address these needs revolve around named entities (e.g., persons, organizations, locations): finding entities, uncovering properties of entities, and discovering relations between them [9].

Online news has become a commodity, available for free from a broad range of sources. To attract readers (and especially readers willing to pay) for their offerings, news distributors are experimenting with new publication models and semantically enhanced news offerings, as witnessed by, e.g., OpenCalais, Guardian, New York Times, Washington Post, etc.

We present SaHaRa, a system for entity-oriented access to news collections. Our goal is to allow users to make better sense of news by embedding it in the linked open data space. We enable users reading an article to learn more about an entity mentioned in the text by offering them a summary of what objects (topics or other entities) it is associated with; given a topic of interest, we identify related entities.

To develop and feed such facilities, we need to detect and normalize named entities in news articles. To discover relationships between topics and entities, we apply techniques from statistical language modeling. In addition, links are formed to entities in DBPedia to facilitate further linkage and discovery of related resources.

An important contribution of this demo is that it marries two traditions: information retrieval as evaluated at, e.g., TREC, and the semantic web; it is important to see the the two communities being connected.

2. RELATED WORK

The task of discovering semantic relations between concepts is core to several Semantic Web tasks such as ontology matching, ontology learning or ontology enrichment. The majority of approaches from the ontology learning community have primarily focused on exploring textual sources for relation learning [4]. Besides these, there are efforts to discover relations from structured data, for example, the DBPedia relation finder [8]. The novelty of SaHaRa is that it relates information from disparate sources, and integrates both textual and structured data.

Entity-oriented search (as opposed to document search) attracts attention from academia and industry. In 2005, an expert finding track was launched at the TREC Enterprise track, where a ranked list of experts had to be returned for a given topic [1]. The INEX XML Entity Ranking track aims to create a test collection for entity retrieval on Wikipedia [5]. The Entity track at TREC (launching this year) evaluates entity-related search tasks on the Web [13]. Many entity-oriented search tasks can effectively be treated as "association finding" between topics (terms) and entities or between entities and entities [2]. State-of-the-art methods are capable of capturing and estimating the strength of these associations by observing the language usage around entities [2, 10, 11]. Commercially available entity-oriented search facilities deal with people, companies, services, locations etc. [7, 6, 14].

News search engines like Google News and EMM Newsexplorer cluster news stories based on textual content. Lifting this to tracking topics and entities yields a new, and, according to news providers, desirable user experience.

3. MAIN CHALLENGES

Complementing entity search with linked open data presents a number of challenges. Entities are not represented directly (as retrievable units such as documents), and we need to identify them "indirectly" through occurrences in documents. The main challenges that had to be addressed in realizing SaHaRa, concern (1) the recognition of entities in documents, (2) the way in which entities are represented, (3) matching topic descriptions and entities, (4) linking entities, and (5) removing noisy data (duplicates, edits, etc).

4. SYSTEM OVERVIEW

SaHaRa provides access to Dutch news provided by ANP, the leading Dutch news agency. SaHaRa provides three main "views" on news; see Figure 1.

- **Search.** Showing matched articles on the left and matched entities on the right.
- **Article.** The news article is shown along with metadata (e.g., category). Entities in the text are linked and are displayed on the right, grouped by entity type. The system offers related news based on associated entities.
- **Entity.** If available, the first section of the Wikipedia page is displayed at the top. Below, the entity's (language) model, i.e., terms strongly associated with the en-

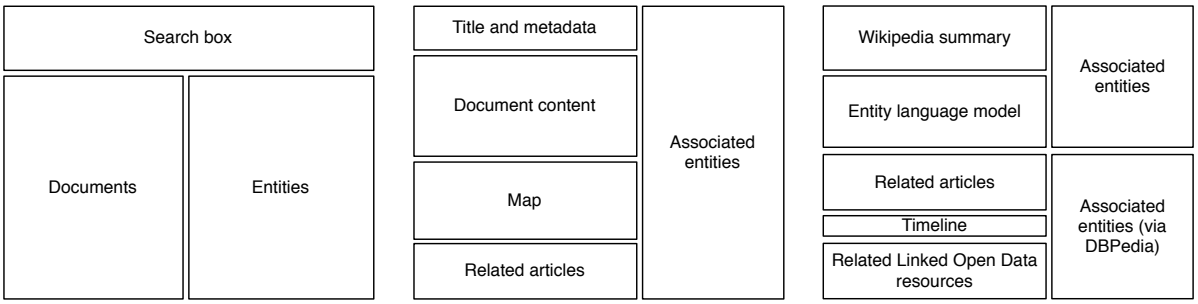

Figure 1: Schematic overview of SaHaRa's views on news. (Left): search. (Middle): article. (Right): entities.

tity and associated entities—both as tagclouds; the size of the text indicates the strength of the association. The page also shows (1) documents about the entity, (2) a graph indicating the frequency of mentions of the entity over time, (3) resources from the Linked Open Data cloud about the entity (e.g., documents from Wikipedia), and (4) related entities via DBPedia.

We sketch the key components that feed these views:

Data Collection and Preprocessing. The current public version of SaHaRa provides access to 152,015 documents from 2007. These articles are published along with some metadata, such as category and author. After filtering out articles without title, internal memos, overviews of news and papers, and resubmissions of the same article, we were left with 93,725 documents.

Named Entity Recognition. A named entity tagged version of the collection was created using the TnT [12] statistical part-of-speech tagger trained for Dutch. Four standard types of entities were detected: person, organization, location, and miscellaneous; in total 288,743 unique entities exist in the public demonstrator.

Entity Modeling and Search. SaHaRa represents entities using language models (i.e., multinomial probability distributions over terms). These models (or profiles) are constructed using the EARS toolkit [3]. On the entity profile page, SaHaRa displays associated terms and entities as a tagcloud; these are terms with the highest weight in the entity model. When searching for a query, SaHaRa ranks entities according to the probability of generating the query, analogously to language modeling for document retrieval.

Linking to DBPedia. To provide background information and for normalization purposes, SaHaRa matches entities in the news with entities in DBPedia. Links to DBPedia are also used for query suggestions. Additionally, if the entity can be unambiguously matched to a DBPedia entry, we display the first section of the corresponding Wikipedia article on its profile; if it exists, an image of the entity is also presented. Finally, we search for associated entities in DBPedia using category information. This enables us to discover related entities beyond the scope of the news collection.

SaHaRa uses the Dutch Wikipedia and an NE recognition system trained on Dutch data; it can be applied to any language for which these two resources are available.

5. DEMONSTRATION

SaHaRa will be demonstrated on Dutch news, but the user interface is available in English as well. Given the focus on entities, the demonstration can easily be appreciated by anyone who understands English. The demonstration will illustrate temporal dimensions, show associations between different types of entities, and highlight how linking across articles and to background sources provides valuable enhancements.

6. SYSTEM REQUIREMENTS

To demonstrate SaHaRa a regular web browser and internet connection are needed. The publicly accessible version of SaHaRa is available at `http://ilps.science.uva.nl/demo/sahara`. The authors will also prepare an offline version of the system.

Acknowledgments. This research was supported by the Netherlands Organization for Scientific Research (640.001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802), the Dutch-Flemish research programme STEVIN (STE-09-12), and an innovation voucher from the Dutch Ministry of Economic Affairs. We are very grateful to ANP for providing a year worth of news data.

7. REFERENCES

[1] P. Bailey et al. Overview of the TREC 2008 enterprise track. In *Working Notes TREC 2008*, 2008.

[2] K. Balog. *People Search in the Enterprise*. PhD thesis, University of Amsterdam, June 2008.

[3] K. Balog. Entity and Association Retrieval System (EARS), 2009. URL: `http://http://code.google.com/p/ears/`.

[4] P. Cimiano. *Ontology Learning and Population from Text*. Springer New York, Inc., 2006.

[5] A. P. de Vries et al. Overview of the INEX 2007 entity ranking track. In *INEX 2007*, pages 245–251. Springer, 2008.

[6] Evri. Search less, understand more, 2009. URL: `http://www.evri.com/`.

[7] Google. Google, 2009. URL: `http://www.google.com/`.

[8] J. Lehmann, J. Schüppel, and S. Auer. Discovering unknown connections - the DBpedia relationship finder. In *CSSW'07*, pages 99–110, 2007.

[9] D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers. *Analyzing Entities and Topics in News Articles Using Statistical Topic Models*. Springer, 2006.

[10] D. Petkova and W. B. Croft. Proximity-based document representation for named entity retrieval. In *CIKM '07*, pages 731–740. ACM, 2007.

[11] H. Raghavan, J. Allan, and A. Mccallum. An exploration of entity models, collective classification and relation description. In *KDD '04*, pages 1–10, 2004.

[12] TnT. Statistical part-of-speech tagging, 1998. URL: `http://www.coli.uni-saarland.de/~thorsten/tnt/`.

[13] TREC. TREC Entity track, 2009. URL: `http://ilps.science.uva.nl/trec-entity/`.

[14] Yahoo! Research. Correlator, 2009. URL: `http://correlator.sandbox.yahoo.net/`.