Adaptive Semantic Publishing

Georgi Georgiev, Borislav Popov, Petya Osenova, Marin Dimitrov

Ontotext AD, Bulgaria
{borislav.popov, georgiev, petya.osenova,
marin.dimitrov}@ontotext.com

Abstract. The paper describes the approach, methodology and main software components of an Adaptive Semantic Publishing Platform for digital medias; applied previously to numerous use cases and publishers like the BBC, Euro-Money and Press Association. The semantic publishing relies on the interaction among the common sense model in ontologies, the world knowledge in Linked Open Data (LOD), the named entity categorization and the set of domain-specific keywords. Hence, the contribution of the related LOD datasets is briefly considered. The adaptive publishing relies on the user's requirements (interests, searches, activities) provided as summaries of articles on selected topics (sports, politics, society, etc.). Also, approaches to gold standard data are presented, which enable the fast and high quality clusterization of numerous information streams per topic.

Keywords: Semantic Publishing, Personalization, Clustering, Ontologies, Linked Open Data, Summarization

Introduction

In recent years Semantic publishing applications get more and more user-oriented in several aspects, among which: customization and re-purpose of data and content reflecting the user needs; focused summaries with respect to user interests; high relevance of the retrieved information and minimal effort in receiving it.

There are various works, exploring the relation between publishing and Linked Open Data. In [4], for example, authors present their idea on a life cycle model (specification, modeling, generation, linking, publication, exploitation) and demonstrate its application within various domains. At the same time, in [3] a DBpedia service has been presented (called DBpedia Spotlight), which automatically annotates text documents with DBpedia URI's using the DBpedia in-house ontology. Similarly, Zemanta[1] provides a plug-in to content creators, which recommends links to relevant content (articles, keywords, tags). Our approach is generally in-line with these ideas and services – domain specific applications, automatic semantic annotation, adding relevant linked content. However, our focus is preferably on: the trade-off between the seman-

[1] http://en.wikipedia.org/wiki/Zemanta

tic knowledge holders (ontologies, linked data) and their language reflection (domain texts), mediated by the linguistic processing pipelines; the adaptive flexibility of the constructed applications and the efficient storage and publishing of large data.

Within Ontotext, examples of mass media, semantic publishing web sites, such as the BBC's sport web[2] and the official web of the London's Olympics 2013, have proven to attract a multi-million user bases. Behind such applications, as revealed by lead engineers at the BBC[3], there lies the complex architecture of the state-of-the-art Semantic and Text Analytics technologies, such as in-house: fast RDF database management system OWLIM[4] and knowledge management platforms KIM[5]; for robust semantic annotation and search, as well as for text analytics applications.

Both platforms are incorporated into numerous successful Semantic Publishing Solutions (including the BBC Sport[6], Press Association[7], Newz[8], EuroMoney[9], Publicis[10] etc.). This paper aims to describe the approach, main software components, information architecture, text analytics and semantic annotation and indexing, used successfully in many solutions for more than 5 years, to build semantic publishing solutions.

Our approach relies on the calibration between the RDF semantic repository OWLIM, the semantic resources in KIM and the optimized Text Analytics techniques including methodologies for fast creation of gold data in the selected domain; focused curation of the automatically analyzed data and the application of advanced machine learning algorithms in data clustering. Thus, the success of our solutions lies in the customization of the advanced semantic technologies in combination with text analytics techniques, tuned to the needs of publishers and adapted to the requested domains.

The Overall Architecture of Semantic Publishing System

Our generalized system, presented on Fig. 1 below, comprises several components, connected in a life cycle. The Content Store on the left side contains the news articles, along with the associated pictures and videos. The textual content of these assets is then passed to the Text Processing Pipeline, implemented in GATE[11]. The pipeline includes various components. The generic components refer to: tokenization, POS tagging, chunking. The Gazetteer refers to named entity recognition, such as Person, Location, Organization, etc. The Rules are written in JAPE[12] style. They

[2] www.bbc.com/sport
[3] www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html
[4] www.ontotext.com/owlim
[5] http://www.ontotext.com/kim
[6] http://www.ontotext.com/publishing
[7] http://www.pressassociation.com/
[8] newz.nl
[9] http://www.euromoney.com/
[10] http://www.publicis.de/
[11] http://gate.ac.uk/
[12] http://gate.ac.uk/sale/tao/splitch8.html#chap:jape

usually cover the relation extraction, such as Person works for Organization; Organization is located in Location, etc. Machine learning component scales up the application when there is manually annotated training data. Semantic indexing refers to URIs, which map the detected entities with real objects in the world. For example, Obama is recognized as Person, but then, more information about him is provided through a link to Obama's DBPedia[13] profile. Geo localization of the articles relies on automatic association with GeoNames[14] map coordinates. This means that the recognized country is assigned with its longitude and latitude information from the interlinked map.

Additionally, some other knowledge is provided, such as capital, currency, etc. In this way the facts are available in interconnected knowledge maps. The ontology abstracts over the text chunks and specifies the categories and relations within linked data. We initially rely on common sense ontology (such as PROTON), which might be further extended with the necessary domain knowledge depending on the required conceptualization granularity. Given the common ontology and the data, the extension is trivial and easy.

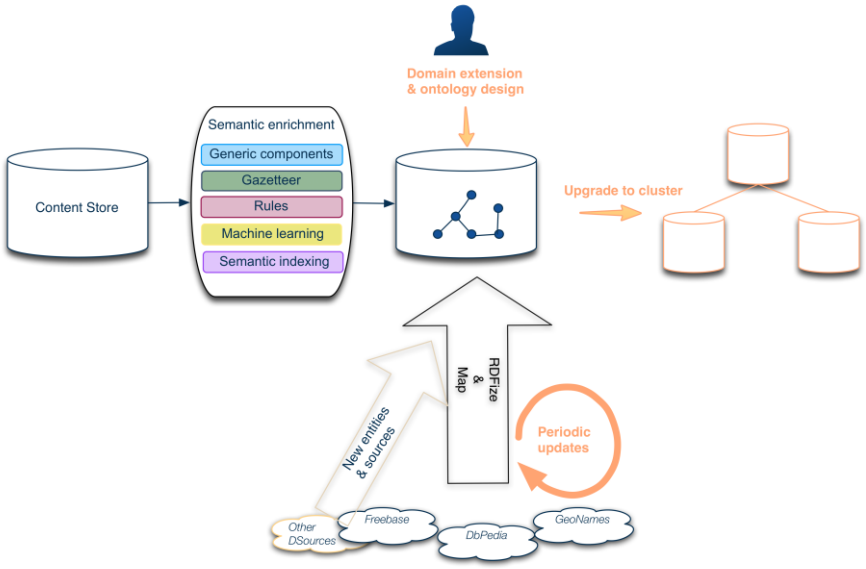

Fig. 1. Generalized Semantic Publishing Architecture.

[13] http://dbpedia.org/About
[14] http://www.geonames.org/

In Fig. 2 below the Basic Information Architecture is shown. The Domain Extension and Ontology design process include modeling of the specific domain (finance, sports, etc.) in domain ontology, connected to the upper-level PROTON ontology[15]. At the same time, the processed data is mapped to Linked Open Data. Linked Open Data also provides updates on named entities.

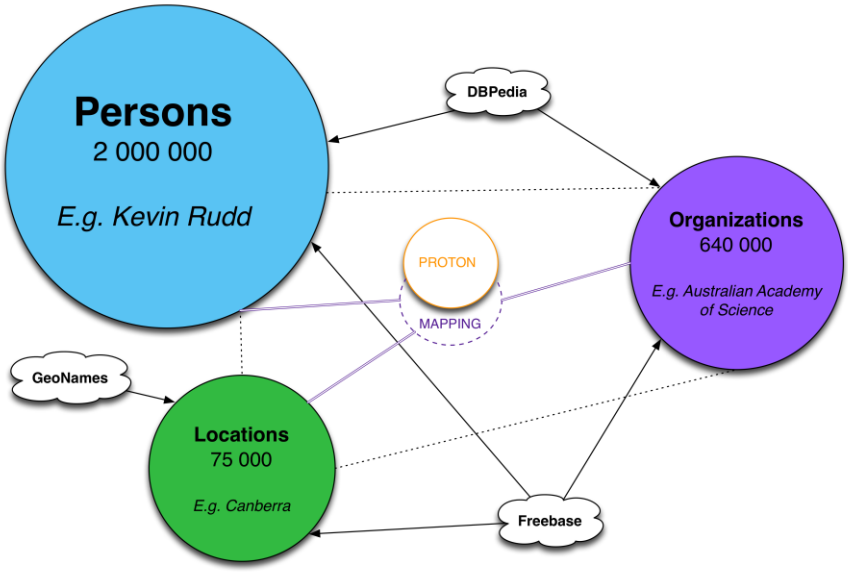

Fig. 2. Basic Information Architecture. The arrowed lines show what types of concepts are present in the specific linked resource (for example, GeoNames have only Locations). The dotted lines show that there might be various relations among Persons, Locations, Organizations (such as, Person lives-in Location). All the instances from linked data are mapped to a common ontology (PROTON).

The interconnected data then goes into clustering models for focused categorization with respect to the task. We work in a hierarchical way of knowledge recognition – first, we try to detect a referent in LOD (i.e., that Obama is a Person with features like birthday, birthplace; LivesIn; is_a president of the USA in the appointed time periods, etc.); if there is no such connection, then we aim at recognizing the more abstract concept in the Ontology (*Person with a Role in Society*, and with other appropriate features); if this information is also not available for some reason, then the named-entity is detected (*Obama is Person*); if this is not possible either, then the most important key phrases are recognized (the terminology in the domain, etc.); last, but not least, topics of documents can be categorized (Sports, Society, Politics,

[15] proton.semanticweb.org

Finance, etc.). In most of our use cases, however, all these types of recognition are available and connected to each other. Thus, clustering of data is based on detailed semantic annotations.

The recognized knowledge pieces (referents, concepts, entities, etc.) can be also cross-related. For example, the categorization of entities in Sports, will provide groups of Persons, involved in Sports activities, Types of Sports, etc. On the other hand, information might be given about most popular sports per country or about Sports careers of Persons in various Organizations.

The Semantic Annotation Solution

The Underlying Datasets.

The OWLIM repository uses the following datasets: Freebase (version jan_9_2012), DBpedia (version 3.8) and subset of Geonames. From Freebase and DBpedia all the people and organizations are integrated. The subset of Geonames includes bigger cities, countries, continents, oceans, seas, US states, selected places in Europe. All these linked open data sets have their own ontologies, which are also part of the repository. All the above-mentioned ontologies are mapped to the common sense upper-level PROTON ontology. The added value of such a mapping is that a high quality reasoning and consistency of the modeled knowledge is ensured.

Since there is duplicated information in the three data sets, which, however, is presented in different ways, a mapping effort has been performed also between Freebase and DBpedia; Geonames and DBpedia. Thus, via the mappings of two bases to DBpedia, Freebase has also its mapping to Geonames.

Gold Standard Annotation and Curation.

The gold standard annotation includes the following steps: understanding the task; preparation of annotation guidelines; manual annotation of some texts with pre-selected tags; training an algorithm over gold data; automatic annotation of big datasets; curation of the processed documents; re-training (if necessary). For more details see [1] and [2].

The curated documents are used in turn as a bigger gold standard corpus for automatic text analysis evaluation, but also for training further machine learning models over the data.

The level of granularity for the creation of Annotation Types and their Features is based on more concrete classes for Person, Organisation, Location of the PROTON Ontology, with extension of classes and properties from linked open resources, such as DBpedia, Freebase and GeoNames. Modularization depends also on the specific task. In the case of media publishing, it detects Person, Organization and Location and maps them to the descriptions in Linked Open Data sources. But it also respects the domains, in which the named entities occur (Sports, Politics, Economics, etc.)

The document curation subsumes three related tasks:

- *Instance disambiguation*. It handles cases, such as ambiguities between various people with the same name or various locations with the same name, or even

person, location and/or organization with the same name. Since such properties are very context-dependent, the possible true candidates are verified by assigning the correct URL from Linked Open Data. For example, the name 'Washington' is ambiguous, since it refers to a *Politician*, an *Actor*, a *US state* and a *US city*.

- *Tagging*. This step is applied only for **Person**, **Organisation** and **Location**. The labels **Person**, **Organisation** or **Location** assigned during the automated annotation stage are verified and new labels are added to specific words or phrases in the text. This step prevents from considering a *Person* as a *Location*, or vice versa. The step is very important, since such mis-tagged entities might not be many, but might be very frequent in the domain. On Fig 2 above the complex relations are given among named entities, such as *Persons*, *Organizations* and *Locations*, etc. and their Linked Open Data mappings.
- *Topic or key words correction*. This step is applied when a document is miscategorized for a topic (for example *Society* instead of *Finance*) or when some phrases are detected which do not belong to the terminology of the domain.

In Fig. 3 below the activities cycle behind the Basic Semantic Annotation Process is shown.

From the perspective of the Agent, it contains the following actions. The reports or articles, provided by the client (the person on the top), are carefully examined by the company specialists (SME – small medium enterprise).

Then some probe annotations are performed for clarifying and defining the required annotation types. Then bigger chunks of data are annotated by annotators while keeping high interannotator agreement and providing checks by a superannotator.

When the data has been automatically annotated via the trained manually produced data, then also the curator is involved.

He/she chooses the correct mappings from a set of available mappings. Very often the set contains ambiguous mappings, which have to be resolved.

He/she also adds new links or deletes some, if needed. Simultaneously, the information from Linked Open Data store is mapped to the semantic annotations (both manual and automatic).

These mappings need a careful curation for achieving a high precision and recall. The person on the right indicates the client, who might examine the annotated data before its linking to the open data and its storage. This involvement is optional.

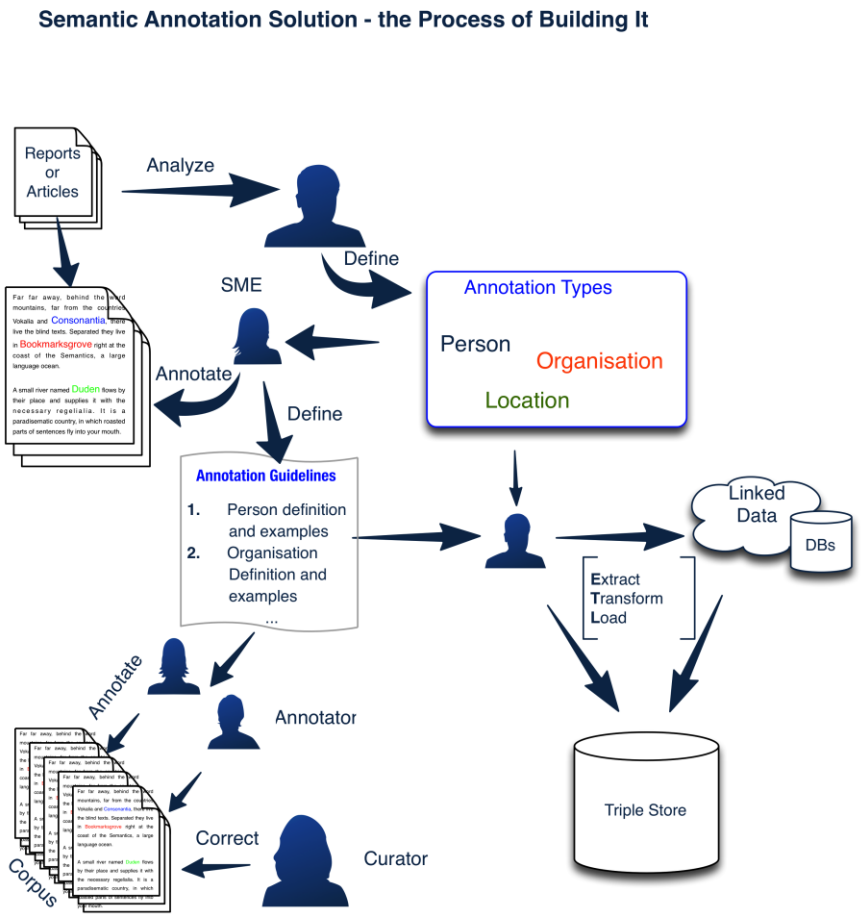

Fig. 3. Basic Semantic Annotation Architecture. The persons indicate human intervention in the automatic process.

Personal Semantic Publishing

The Semantic Publishing Platform provides the following adaptive instruments: parallel streams of excerpts from topic-related articles (trending); personalized user area (profile); search by topic-related keywords and focused summary. Let us consider each service separately.

Trending provides parallel streams of articles on various topics. The user can get oriented within the variety of topics, can select the ones of interest to him and explore them further (see Fig. 4).

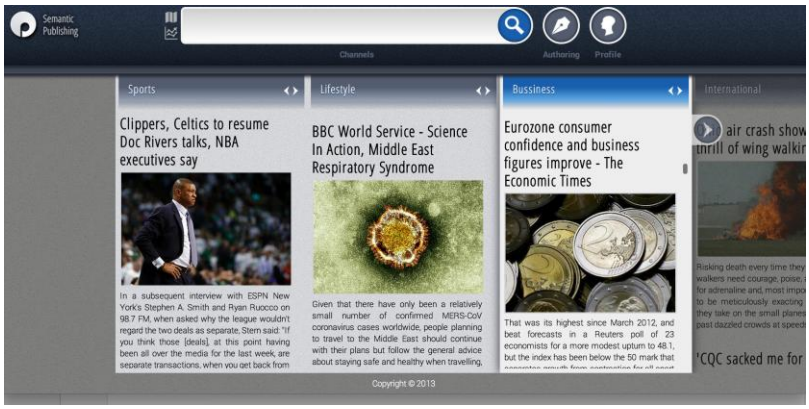

Fig. 4. Trending: parallel news streams.

The personalized user area allows the user to mark the articles and summaries that are of interest to him, to tag them with pre-selected tags and to store them for further usage. Additionally, the user can see related articles to the selected topic (see Fig. 5).

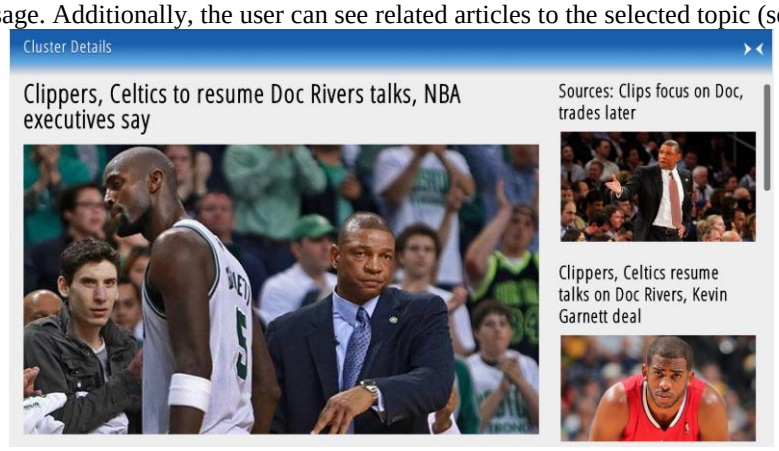

Fig. 5. Showing similar content to the chosen article.

The user can search information through sequencing of as many keywords and concepts as necessary for constraining the requested topic. Also, an autocomplete search is added as a facility (see Fig. 6).

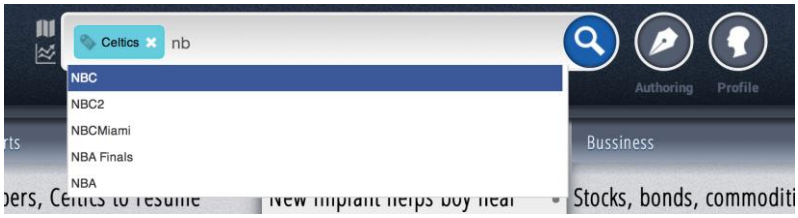

Fig. 6. Showing autocomplete options.

The user can get a summary of the topic, customized from related articles and built on the clustered data. Additionally, he/she receives the most frequent named entities and keywords in a rated list (see Fig. 7).

Fig. 7. Showing summary and related keywords.

Conclusions

The current work reveals a platform and methodology for development of Adaptive Semantic Publishing solution, applied previously in many use cases such as the BBC Sport web site. A concrete solution is described in terms of methodology and main software components and is publicly available as demonstration software.

The main user interface components: faceted search, trending, term and word search as well as channel's customization are also described with examples. The implementation of the user interfaces for personalization/profile and authoring are main areas of future work.

References

1. Kiryakov et. al 2003: Atanas Kiryakov, Borislav Popov, Damyan Ognyanoff, Dimitar Manov, Angel Kirilov, Miroslav Goranov. Semantic Annotation, Indexing, and Retrieval. 2nd International Semantic Web Conference (ISWC2003), 20-23 October 2003, Florida, USA. LNAI Vol. 2870, pp. 484-499, Springer-Verlag Berlin Heidelberg 2003.
2. Popov et al. 2003: Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, Miroslav Goranov. KIM – Semantic Annotation Platform,
3. 2nd International Semantic Web Conference (ISWC2003), 20-23 October 2003, Florida, USA. LNAI Vol. 2870, pp. 834-849, Springer-Verlag Berlin Heidelberg 2003.

4. Mendes et. al 2011: Pablo N. Mendes, Max Jakob, Andres Garcia-Silva and Christian Bizer. DBpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems, pp. 1-8, ACM, New York, NY, USA.
5. Villazon-Terrazas et. al 2012: Boris Villazon-Terrazas, Daniel Vila-Suero, Daniel Garijo, Luis M. Vilches-Blazquez, Maria Poveda-Villalon, Jose Mora, Oscar Corcho, and Asuncion Gomez-Perez. Publishing Linked Data - There is no One-Size-Fits-All Formula. In: Proceedings of the European Data Forum 2012, Copenague, Dinamarca.