Semantically Enriching Content Using OpenCalais

Marius-Gabriel BUTUC

Abstract—One of the key challenges of the Semantic Web is how to go from today’s unstructured web to a web rich with semantic information. Following the bottom up approach, OpenCalais is an automated system intended to annotate data and information, allowing us to gradually build semantically enabled systems. By semantically enriching the published content, we help users enjoy their on-line experiences and reduce the frustration of dealing with voluminous amounts of information that is incoherently organized and often irrelevant to a particular person’s need.

Index Terms—API, Linked Data, resource management, web service.

I. INTRODUCTION

CALAIS identifies and automatically tags entities (e.g., people, places, companies, geographies), facts (e.g., relationships) and events (e.g., things that happened) in text. It then fabricates connections between those entities and significant data sets, media files, and similar entries using open data from sources like Wikipedia, DBpedia, GeoNames, the Internet Movie Database (IMDB), Shopping.com, Reuters.com, etc.

With the original goal - help developers, bloggers and publishers automatically tag their content to improve search and navigation, and enable new reader engagement features - the team behind Calais has developed a tool to automate time consuming content operations and increase productivity.

This paper will look into the possibilities of semantically enriching published content [1] using Calais.

II. CALAIS AND THE OPENCALAIS WEB SERVICE

Calais is described by its developers as a “big initiative with a lot of components” [2], yet at its core lays the OpenCalais web service. The web service is an API that accepts unstructured text, processes it using Natural Language Processing (NLP) and Machine Learning (ML) algorithms, returns RDF-formatted entities, facts and events and is used to fuel the applications that build the Calais initiative.

An overview of the Calais applications and tools is depicted in Fig. 1.

Recently nominated one of the Top 10 Semantic Web products of 2009 [3], OpenCalais uses semantic technology and natural language processing to analyze text and add metadata by drawing out entities from documents, blog posts, news stories, etc. In some cases [4], this type of data can identify or help identify relationships between people, businesses, etc.

A. How does it work?

The basic interaction delivered by the OpenCalais API (Figure 2) can be summed up in the following steps:
1) First of all, we need an API key – a string that uniquely identifies the specific instances of the application that uses the service.

2) After obtaining the key, we can submit content to the Calais Web service. The methods supported by the OpenCalais Web Service API are: SOAP, REST and HTTP Traffic Compression.

3) Calais tags each person, place, fact and event in the content, making it machine-readable and interoperable on the Web.

4) Each piece of content – and each entity or event in that content – is assigned a unique identifier (a document ID and many URIs) that ties back to the Linked Data Cloud.

5) We can use the metadata returned by Calais (tags, document IDs and URIs) in our application, e.g., we can use OpenCalais to scan a set of Web pages and build a local RDF store for querying and display.

B. LinkedData Entities

As of March 2009, Calais is officially part of the Linked Open Data (LOD) Cloud (Figure 3).

The Calais initiative has exposed its data via Linked Data endpoints. When Calais extracts an entity from a given text it also returns a dereferenceable entity URI. All the naming conventions for entities, facts, events and their corresponding attributes are rationalized and the schema is published on the official site as RDFS.

There are several classes of entities that are considered meaningful to Calais and they include Currency, MarketIndex, Movie, MusicGroup, MusicAlbum, OperatingSystem, ProgrammingLanguage and many others. The breadth and depth of information provided in each URI varies based on how well Calais can unambiguously identify this entity, automatically using its specific NLP algorithms.

As an example, “Calais” is an ambiguous name: we have both Calais – the semantic tool and Calais – the city. Even as a city name Calais is ambiguous, because there is one in the North of France and several in the USA. The URI of the ambiguous city “Calais” is:

http://d.opencalais.com/genericHasher-1/82b41c38-0b5e-301f-8a84-839d4d78e78e.html

If we try to dereference this URI, we’ll get disambiguation options, each leading to another resource known to the API that can disambiguate “Calais”. If we follow any of those resources, we will find useful information such as geographical coordinates and links to other Linked Data or Web assets.

Currently the list of disambiguated entities is limited to Company, Product (Electronics) and Geographies: City, Country and ProvinceOrState. Company, for example, is the richest of endpoints providing rich information, such as ticker symbol, officers and directors, corporate website, industry codes, etc.

C. Interpreting the API Response

The Calais response can be easily understood [7] from RDF, JSON, microformats [8], [9] – a web-based approach to semantic markup that seeks to re-use existing (X)HTML tags to convey metadata and other attributes – or a simple HTML format. Another example of the implementation of the LOD principles is given when we look into the RDF representation for Moscow, Russia. This entity is linked via owl:sameAs to

- http://dbpedia.org/resource/Moscow
- http://sws.geonames.org/524901/
- http://rdf.freebase.co/ns/guid.9202a8c04000641f800000000002636c

III. EXAMPLES OF APPLICATIONS THAT USE OPENCALAIS

A. Calais Viewer

Calais Viewer [10] is the best way to get a quick glimpse of OpenCalais. We can manually try the web service by typing in or pasting text into the viewer box, without needing an API key. For example, we used a Wikinews teaser about Barack Obama and got the results depicted in Figure 4. It identifies
the correct topic of the article – Politics, the City that the article is about – Washington, United States, two Person entities: Barack Obama and George W. Bush etc.

B. Tagaroo

As we saw in Figure 1, Calais provides Content Management Tools for Drupal and Wordpress. Tagaroo [11] is one of those tools aimed to make any WordPress blog better for the publisher, for the readers and more accessible to search engines. Similar to Zemanta [12], Tagaroo analyzes the text in our post and suggests intelligent tags for the things and events we’re writing about. Before using it, we need to have a valid API key (Figure 5).

C. SemanticProxy

Following the standard for publishing Linked Data on the Web [13], the Calais team developed SemanticProxy [14] to translate the content of any URL on the web to its semantic representation in RDF, JSON, microformats or HTML (Figure 8). It is optimized for performance on 30 of the world’s largest English-language news sites, but it also works well with other sites like in Figure 7. Intended for the “little semantic machines love to come by and spend a few highquality milliseconds” [14], it is meant to leverage both current and future semantic applications.
IV. CONCLUSION

OpenCalais currently only supports content in English, French and Spanish and it tries to identify automatically among the given languages by applying a Language Identification module before processing the text for entities, events and facts. Supporting more languages could be a good direction for development, opening Calais even more. The web service is free for both commercial and non-commercial use. The current limitations are 50,000 transactions per day with a frequency of maximum 4 transactions per second while an average transaction is expected to take from 0.5 to 1 second.

Using a semantic-based tool, the users are able to easily automate content operations, increase productivity and cut costs. Since the Semantic Web goal is not to build a web of data, but to enrich lives through access to information [15], we can now approach the bottom up paradigm [16], enhancing the value of our content and improving the user experience.

As stated in Section II-A, we can use OpenCalais, or even better SemanticProxy, to scan a set of Web pages and build a local RDF store for querying and display.

Another application could consist in a visualization tool that could make OpenCalais even more powerful. For example it might be interesting for visualization tools like Muckety [17] or NNDB Mapper [18] and to quickly "see" relationships that might go unnoticed without tools like OpenCalais.

We intend to experiment the use of OpenCalais for enhancing the content of the academic social Web applications (e.g., blogs, wikis) in the context of the semantic e-learning.

REFERENCES