Entity Extraction

Shivangi Saxena
ss4733@columbia.edu

OpenCalais provides a simple and straightforward way of analyzing documents and websites. It is easy enough for someone with limited knowledge of coding to use and thus greatly broadens its user base. If necessary, services like *semanticproxy.com* can be used to perform the content scraping, since the OpenCalais API itself requires scraping to be done manually.

Having said that, OpenCalais is not very accurate and also tends to miss out on a lot of entities. OpenCalais works best when dealing with locations. In the sample documents I used, it recognized all the places given to it, whether they were countries or state (thanks to the huge dataset of Linking Open Data cloud).

It also worked well in detecting big company names, like Google and Facebook. It did, however, face issues in detecting the names of companies that were new/small-scale or start-ups. This is to be expected, as newer/ smaller companies may not have as much data about them in the LOD cloud. It could also not detect abbreviations, such as in document 1, where it detected "*The Royal Society for Public Health*" but not "*RSPH*".

Of the many entities detected, the ones of type "*IndustryTerm"* tended to be trivial most of the times, detecting terms like "food" and "electricity", which didn't really require hyperlinks.

OpenCalais also had a tough time detecting references to entities that had metaphors. This can be seen in document 3, where Greenpeace is referred to as "the green group", an entity that Calais detects as a reference to the color green.

The given hyperlinking strategy works in most cases, unless a person/object does not have a Wikipedia page, which may or may not be a problem depending on the kind of documents that are being analyzed. When analyzing major news articles, this tends to be rare. Similar to the case given in the Calais documentation for RDF documents, when working with ambiguous entities, Wikipedia too shows a page with disambiguation options. This can be seen in document 5, where clicking on "Cleburne County" lets us choose between the ones in Alabama and Arkansas.

	Entity References	Detected References	Correct References
Document 1	17	12	7
Document 2	18	10	5
Document 3	19	12	8
Document 4	8	6	2
Document 5	13	13	10
Total:	**90**	**53**	**32**

As seen above, OpenCalais has given about a 59% accuracy in detecting entities and of these, 60% tended to be correct links (i.e., either pointing to the RDF document or the Wikipedia page).

Although no NLP – system is ever going to be a 100% accurate, I believe it is safe to assume that with greater expansion in the data set of LOD, the accuracy of OpenCalais would greatly increase.