23^{rd} of April, 2008
Guest lecture at Institute of Technology, Tallaght, Ireland

Ontology-based Text Processing for Personal Information Management

Alexander Troussov, PhD. IBM Dublin Center for Advanced Studies Chief Scientist

ABSTRACT:

This talk is based on IBM's experiences in the Nepomuk project. Nepomuk aims to build a socio-semantic desktop based on semantic web technologies centred around a personal information management ontology, PIMO, which represents each user's perspective on his/her data, the concepts they refer to and how they all relate to each other and the real world.

However, for most applications to benefit from the use of semantic web technologies, the user needs to manually annotate his/her data (emails, documents etc) with information (from his/her own perspective) as to what they are about, how they relate to each other and to topics/concepts in the ontology.

Fortunately, this need not be the case. PIMO is a kind of ontology, and can be used within text analysis to facilitate concept detection, generalisation, disambiguation automatic content labelling and much more.

But how?

Firstly we must consider how most approaches tend to treat lexical entries in ontologies. Typically this follows one of two approaches

A) as labels which correspond to a "name" for the item/concept represented by a node in the ontology,
    a. But how to find them in real texts where these labels can be inflected (plural/derivational/etc) or in the case of multiword label there could be intervening tokens, or both?
B) as a commonly agreed vocabulary, and as synonyms
    a. But what if an individual has a different interpretation for the same concept? For example, mentioning of The Statue of Liberty means NY for me, but freedom/independence to others.

Secondly, we need to consider how to use this for text processing. We introduce a new approach using Linguistic Light Annotation, which we use to describe the lexical layer of an ontology and a Linguistic Light Scanner, which uses these annotations to detect variant forms of the lexical realisation of a term mentioned in text.

Finally, we build a semantic model of the text as function on nodes of a graph which represents a user's world view (their PIMO) this shows how text is relevant to each of the concepts discovered. We process this function taking into account empirics; this allows

us to differentiate between semantic models of coherent and cohesive texts as compared to models of random lists of words.

This processing is based on a graphmining technique, created in IBM Dublin (http://alphaworks.ibm.com/tech/galaxy), similar to “spread-of-activation”. In our implementation this technique can be used for soft clustering and fuzzy inferencing. Application of this technique to semantic models of texts allows us to make key concepts prominent (even if they are not explicitly mentioned in text) and provides context sensitive term disambiguation on the fly.

Alexander Troussov, PhD. IBM Dublin Center for Advanced Studies Chief Scientist

IBM Center for Advanced Studies: http://www.ibm.com/ibm/cas/sites/dublin/

Galaxy Project: http://www.alphaworks.ibm.com/tech/galaxy

IBM Internal HomePage: http://w3.ibm.com/bluepages/simpleSearch.wss?searchBy=name&searchFor=troussov

Internet mail: ATrousso@ie.ibm.com   Office: +353-1-815 1906 Mobile: 087-9155058 Fax: +353-1-815 3537