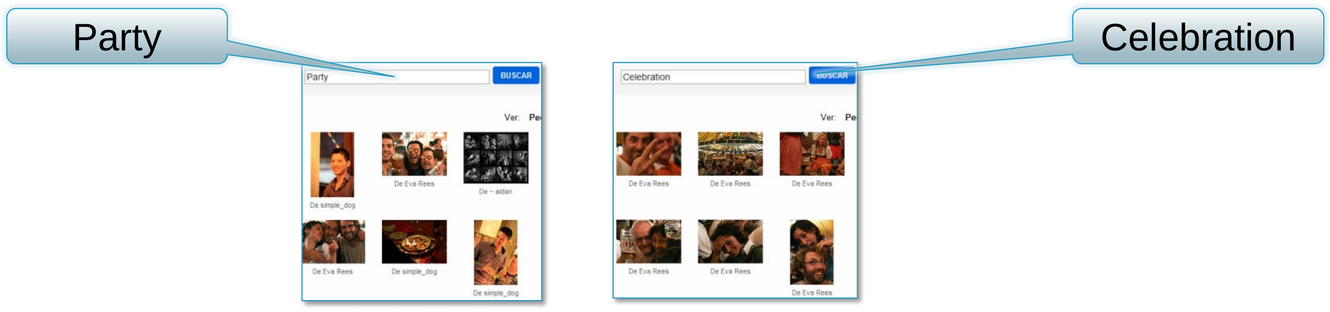
Associating Semantics to Multilingual Tags in Folksonomies

Andrés García-Silva, Jorge Gracia, and Oscar Corcho,
{hgarcia,jgracia,ocorcho@fi.upm.es}
Universidad Politécnica de Madrid, Spain

Abstract: Tagging systems are nowadays a common feature in web sites where user-generated content plays an important role. However, the lack of semantics and multilinguality hamper information retrieval process based on folksonomies. In this paper we propose an approach to bring semantics to multilingual folksonomies. This approach includes a sense disambiguation activity and takes advantage from knowledge generated by the masses in the form of articles, redirection and disambiguation links, and translations in Wikipedia. We use Dbpedia as semantic resource to define the tag meanings.

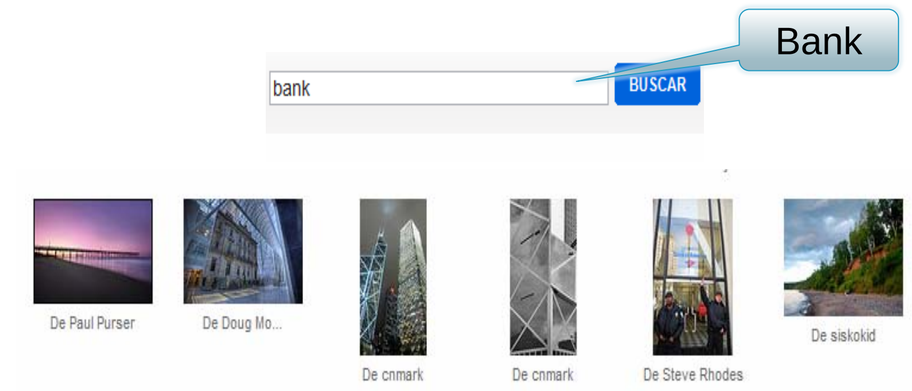
Tagging systems are not aware of:

1) possibly related tags due to relations such as **synonyms**, **broader-than**, **narrower-than**, and **spelling variation**,



Systems ignore *resources tagged with morphological variations or synonyms of that tag, as well as the resources tagged with more generic or more specific tags*

2) the use of **ambiguous** tags.



*When searching with **polysemous** tags, all the resources tagged with that tag are retrieved without taking into account the tag sense the user was looking for (e.g bank retrieves financial institutions and river edges)*

Multilingual Sense Repository

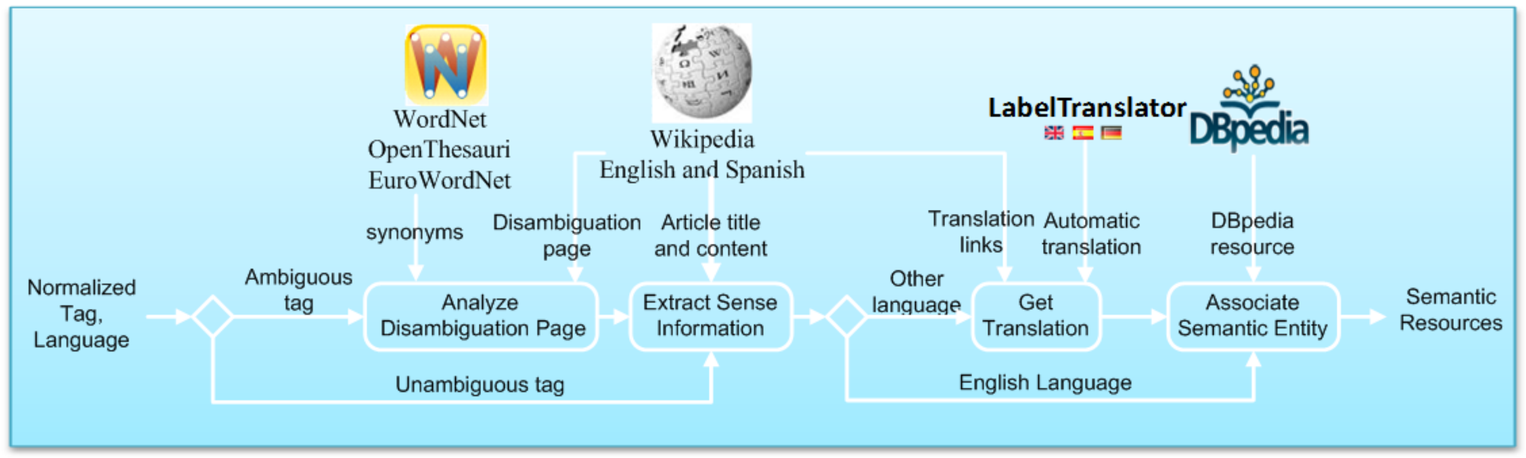
Normalized versions of the tags are related to the set of opossible meanings defined by DBpedia resources

1) Article **URLs** as **sense identiers**, and **article words** along with their **frequency** as **keywords** associated with the **sense**

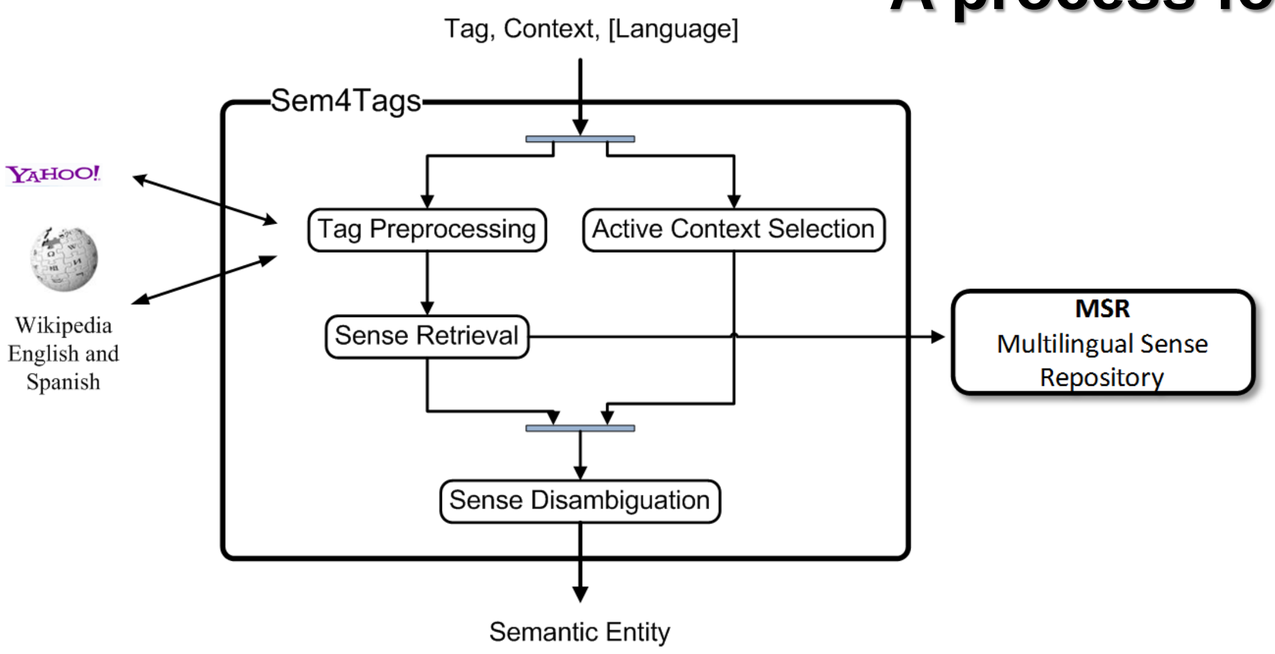
2) Articles listed in **disambiguation pages** as **possible senses** for ambiguous words,

3) Explicit translations among articles to link senses in languages diferent from English to English senses

4) DBpedia resources to define formally each sense.



A process for the Association of Semantics to tags



Vector space model to represent the senses and the tag context.

The vector components are the set of most frequent terms appearing in the Wikipedia pages related to the candidate senses.

In the case of the **sense vectors** the values of these components are calculated using **TF-IDF.**

In the case of the **tag context vector** the values of these components are 1 or
0 whether the corresponding term appears in the tag context or not.

We compare the **tag context vector** against each **sense vector** using the **cosine**.

Experimental results

Test data: 759 photos annotated with 12.4 tags in avarage

Baseline: English and Spanish URI to find Dbpedia resources
http://en.wikipedia.org/wiki/tag
http://es.wikipedia.org/wiki/tag

Coverage		
Approach\Language	English	Spanish
Baseline	51%	32%
Sem4Tags	**83**%	**89**%
Accuracy		
Approach\Language	English	Spanish
Baseline	79%	79%
Sem4Tags	81%	80%
Sem4Tags & Active Context	**86**%	**85**%