Associating Semantics to Multilingual Tags in Folksonomies (Poster)

Andrés García-Silva
Ontology Engineering Group
Universidad Politécnica de Madrid
hgarcia@fi.upm.es

Jorge Gracia
Ontology Engineering Group
Universidad Politécnica de Madrid
jgracia@fi.upm.es

Oscar Corcho
Ontology Engineering Group
Universidad Politécnica de Madrid
ocorcho@fi.upm.es

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[2] as semantic resource to define the tag meanings.

1. INTRODUCTION

The term folksonomy is normally used to refer to the classification schemes that emerge from the tagging activity of a user community. Hence folksonomies represent consensual knowledge, but they are still affected by the lack of semantics. Tagging systems are not aware of: 1) possibly related tags due to relations such as *synonyms, broader-than, narrower-than*, and *spelling variation*, or 2) the use of ambiguous tags.

Despite the fact that tagging systems are web applications with a world wide scope and thus reaching users with multiple languages, semantics of multilingual tags has not been researched. We propose a novel solution for the association of semantics to multilingual tags. Our contribution is twofold: 1) a multilingual sense repository, initially for English and Spanish languages, and 2) Sem4Tags a process for the association of semantics to multilingual tags.

2. RELATED WORK

The problem of identifying tag semantics in Folksonomies has been addressed by researchers in two complementary ways: 1) by identifying groups of related tags using clustering techniques in the hope of such grouping expose the meaning of the tags [6], and 2) by relating Folksonomies with ontologies [1]. In addition, the semantics of relatedness measures among tags has been studied in [3]. However we have not found research works addressing multilingual tags.

3. MULTILINGUAL SENSE REPOSITORY

Inspired by the Tagora sense repository[1] we designed MSR, a multilingual sense repository for English and Spanish based on Wikipedia and DBpedia information. MSR uses: 1) article URLs as sense identifiers, 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) the explicit translations among articles to link senses in languages different from English to English senses, and 4) DBpedia resources[2] to define formally each sense. For each tag to be analyzed the population process is carry out:

Create disambiguation list: First, the list of candidate senses is created. We look for a disambiguation page related to the tag. If this page exists then we extract the possible meanings. Otherwise, we look for a content page related to the tag.

Extract sense information: Then, for each candidate sense we extract the keywords and their frequency from the corresponding article.

Get translations: In addition, for tags in languages different than English, we look for English translations in Wikipedia and using the LabelTranslator tool[3].

Associate semantic entities: Finally, we extract from DBpedia the resources related to the candidate senses. English and Spanish Wikipedia articles are linked to DBpedia resources by means of the `page`[4] and the `wikipage-es`[5] relations. In case the `wikipage-es` relation does not exists for an Spanish article, we use the translation found in the previous activity and use the `page` relation.

4. SEM4TAGS: A PROCESS FOR THE ASSOCIATION OF SEMANTICS TO MULTILINGUAL TAGS

We designed Sem4Tags, a process aiming at associating tags with semantic resources relying on MSR. The input is a tag, its context, and optionally the language of the tag. As context we use the set of user tags co-occurring when

[1] http://tagora.ecs.soton.ac.uk
[2] http://dbpedia.org/
[3] http://neon-toolkit.org/wiki/LabelTranslator
[4] `http://xmlns.com/foaf/0.1/page`
[5] `http://dbpedia.org/property/wikipage-es`

annotating a resource. The output is a DBpedia resource representing the intended meaning of the input tag. The Sem4Tags process includes the following activities:

Preprocessing: The tag is preprocessed to find a normalized representation based on Wikipedia article titles. We benefit from Wikipedia redirection pages when the tag has been considered as an alternative to an article title. In addition, we modify morphologically the tag according to the article title notation. Finally, if after those modifications we have not found a Wikipedia article, we use the Yahoo! spelling service[6] to find an alternative representation.

Active Context Selection: The context is filtered to get rid of tags that can affect the disambiguation activity. The active context contains the set of most highly semantically related tags to the input tag according to a web-based relatedness measure[5].

Sense Retrieval: We select from MSR the set of candidate senses for the tag. We query MSR using the tag normalized version. If the tag is ambiguous the output of this activity is a set of senses. Otherwise, the output is a unique sense.

Disambiguation activity: This activity select the most probable sense for a tag from a set of senses. The idea is that the tag and its context can be compared against each one of the senses measuring the overlapping of the terms in the context with the terms in the Wikipedia pages related to the senses. We use the vector space model to represent the senses and the tag context [4]. 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. Then we compare the tag context vector against each sense vector using the cosine. Finally we choose the sense vector most similar to the input tag as the one representing the intended meaning of the tag.

5. EXPERIMENT

To evaluate Sem4Tags we carried out an experiment using data extracted from Flickr. We gathered 759 photos tagged with tourist cities in Spain (e.g., Barcelona, Ibiza, etc.). On average those photos were annotated with 12.4 tags.

Our baseline attempts to associate directly tags with DBpedia resources. For doing this we create an URI of the form `http://en.wikipedia.org/wiki/tag` for English tags and of the form `http://es.wikipedia.org/wiki/tag` for Spanish tags. Then we query DBpedia for the resource directly related to that URI. For each one of the 2318 tagging activities (i.e., triples of the form $\langle user, tag, photo \rangle$) we run the baseline, Sem4Tags without selecting the active context, and Sem4Tags selecting the Active context. The semantic associations between tags and DBpedia resources where evaluated by 14 users. For the 15% of tagging activities the evaluators were not able to identify the meaning. For the rest of tagging activites the results are shown in table 1.

Table 1: Coverage and accuracy of the analyzed approaches

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

6. CONCLUSIONS

In terms of coverage Sem4Tags is clearly superior to the baseline. This high coverage is due to: 1) the preprocessing activity where tags are normalized, and 2) the amount of information in MSR, specifically the information about the possible meaning of tags. On the contrary, the baseline approach has that low coverage because tags are directly related to Wikipedia content pages, and therefore ambiguous tags, lacking a default article in Wikipedia, are not processed. With respect to accuracy Sem4Tags with Active Context presents the highest value. The use of active context allows us to increase the accuracy in both languages with respect to Sem4Tags. On the other hand, the accuracy of the baseline is very similar to the achieved by Sem4Tags. This fact can suggest that most of the tags are used in the most frequent meaning presented in Wikipedia.

[6] `http://developer.yahoo.com/search/boss/`

7. ACKNOWLEDGMENTS

This work is supported by the GeoBuddies (TSI2007-65677C02) and España Virtual (ALT0317) projects, and the FPI grant (BES-2008-007622).

8. REFERENCES

[1] Angeletou, S., Sabou, M., Motta, E.: Semantically Enriching Folksonomies with FLOR. In 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008). Tenerife, Spain (2008)

[2] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 7, 154-165 (2009)

[3] Cattuto, C., Benz, D., Hotho, A., and Stumme, G.: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In Proceedings of the 7th international Conference on the Semantic Web, Karlsruhe, Germany (2008)

[4] García-Silva, A., Szomszor, M., Alani, H., Corcho, O.: Preliminary Results in Tag Disambiguation using DBpedia. In 1st International Workshop in Collective Knowledge Capturing and Representation (CKCaR09). California, USA (2009).

[5] Gracia, J., Mena, E.: Web-based measure of semantic relatedness. In Proc. of 9th International Conference on Web Information Systems Engineering (WISE 08), Auckland, New Zealand.(2008).

[6] Mika, P.: Ontologies are us: A unified model of social networks and semantics. Journal of Web Semantics 5(1), 5–15 (2007)