Semantic Grounding of Cross-Lingual Folksonomies

Andrés García-Silva[1], Cesar Montaña[1], Iván Cantador[2], and Oscar Corcho[1]

[1] Ontology Engineering Group,
Facultad de Informática, Universidad Politécnica de Madrid, Spain
{hgarcia,cmontana,ocorcho}@fi.upm.es,
[2] Information Retrieval Group,
Escuela Politécnica Superior, Universidad Autonoma de Madrid, Spain
ivan.cantador@uam.es

Abstract. Folksonomies emerge as the result from the free tagging activity of a large number of users over a variety of resources. Though they can be considered as a valuable resource for the classification and exchange of information, several problems have been already identified when it comes to understanding the meaning of tags appearing in folksonomies, mainly related to the appearance of synonymous and ambiguous tags, and specifically in the context of multilinguality. In this paper we evaluate our approach for the semantic grounding of tags to define formally the meaning of multilingual tags. We associate tags with DBpedia resources, according to the context where they were used, by measuring the overlap of terms between the tag context and terms in the Wikipedia articles defining the tag meaning, and we benefit from user translations in Wikipedia to deal with multilingual tags and senses. A conducted experiment shows that the proposed approach achieves high precision for contexts that contain tags in English and Spanish.

1 Introduction

Web 2.0 systems such as *blogs, wikis, multimedia sharing sites*, and *social networking sites* facilitate the creation of user generated content which can be provided in various forms. Among the formats in which user-generated content can be created, tagging has become a popular practice as a lightweight mean to classify and exchange information. Users can assign tags to a wide range of resources including photos, products, urls, and publications among others.

Folksonomies (i.e., the classification scheme emerging from tagging systems) may be considered as a valuable knowledge resource, [11] reports that after some user annotations of a URL, the set of tags used in the annotations tends to stabilize leading to a set of most common used tags to annotate a URL. Similarly, in [16] authors reported the identification of a vocabulary, in terms of tags, among related users in Flickr.

However, folksonomies are affected by the lack of semantics associated with tags [1, 5, 20]. That is, systems are not aware of the use of misspelled tags as

well as synonyms, acronyms, and ambiguous tags. Misspellings, synonyms and acronyms cause that not all the relevant resources are retrieved in user queries, while ambiguous tags cause that irrelevant resources are presented to the user.

Besides, as it happens with the rest of web content, tags are available in multiple languages. We have not found any general statistics about multilinguality in folksonomies though we can cite the results of a recent report on Internet statistics[3], as of December 2009, English speaking users of Internet were only 27.7% of the total. Thus, resources are hidden behind the language barrier created by the tag language. For instance, images tagged with tags in Spanish cannot be retrieved by users who do not know that language. An experiment reported herein shows that in some domains (e.g., Tourism) people use more than one language to tag their pictures. In the analyzed data set we found that 53% of pictures were tagged with tags in English and Spanish.

Some proposals [3, 10, 17, 13] tackle the lacks of semantics associated with tags by clustering them, in the hope of such grouping exposes the meaning of the tags. The clusters are created according to a defined relation among tags, usually relying on a definition of tag similarity [6]. On the other hand, other authors [1, 5, 20] address this problem by relating folksonomies to ontologies. We expect that semantically enriched folksonomies will lead to better information retrieval processes and will be used as knowledge source. In this context, precision can be improved due to the fact that tags are related with their meaning and thus the system can use this semantic information to deliver more accurate results. Semantic relations among tags can be used to carry out query expansion processes [18] so that a richer result set can be presented to the user.

However, the multilinguality of tags has not been addressed by any of the aforementioned approaches. As a matter of fact, those approaches using ontologies are limited to the natural language in which ontology is written, and currently most of the ontologies are written in English. Aiming at overcoming the tag language barrier, we have generated a multilingual sense repository based on Wikipedia and DBpedia [4], and propose an approach (Sem4Tags) for the semantic grounding of multilingual tags in folksonomies that exploits that multilingual sense repository.

The created sense repository contains terms associated with one or more senses in the term language. For each sense we have the keywords appearing in the corresponding Wikipedia article along with their frequency. In addition, senses in Spanish and English are related to DBpedia resources. We chose Wikipedia and DBPedia due to its large coverage and multilinguality support in contrast to some domain ontologies. Wikipedia evolves continuously with user contributions in different languages, while DBPedia represents in RDF a snapshot of some part of the Wikipedia information. Although DBpedia's main language is English, other languages are represented and linked to English resources.

The process for the semantic grounding starts by preprocessing the tag and its context. Then it carries out a sense disambiguation activity where the most probable sense, among the set of candidate senses, is selected according to the

[3] `http://www.internetworldstats.com/stats7.htm`

context. The result of this process is that tags in different languages referring to the same concept are associated with a unique semantic resource. For instance, the *NYC*, and *nueva york*[4] tags used in the sense of New York City, can be related to the semantic entity *New_York_City*[5]

The rest of the paper is structured as follows. Section 2 introduces related work showing how our proposal differs from the others. Section 3 presents our proposed process for the semantic grounding of multilingual tags, and for the generation of our multilingual sense repository. Section 4 describes the experimental evaluation setup. In Section 5 we discuss the evaluation results, which serve as the conclusions of our study. Finally, section 6 presents the future works.

2 Related Work

Folksonomies have been subject of study by researchers in last years. From the semantic point of view two different problems affecting folksonomies have been addressed. The First problem is that **folksonomies lack a uniform representation to facilitate their sharing and reuse**. Some Web 2.0 applications provide APIs to export their folksonomies. However, they do it in proprietary formats. To overcome this problem, ontologies have been proposed to model the tagging activities in folksonomies, with semantic concepts to represent users, tags, resources, etc. A survey in this respect is presented in [15].

The second problem is the **lack of formal and explicit semantic of tags**. Researchers have proposed techniques to discover tag semantics in folksonomies. We have identified three kinds of approaches in this respect. 1) clustering approaches [3, 17, 13] aiming at finding groups of tags relying on relatedness measures among them [6]. 2) Ontology-based approaches [1, 5, 20] whose goal is to associate ontology entities to tags. The ontology can be a domain ontology, or a set of ontologies retrieved by a semantic search engine. Finally, 3) Hybrid approaches [10] use a mix of clustering techniques and ontologies. A survey in this respect is presented in [8]. We want to note that none of the aforementioned approaches deals with multilingual tags explicitly.

Our research work aims at resolving the lack of semantics by grounding tags with semantic entities, unlike [3, 10, 17]. In addition, while we use DBpedia as semantic resource, the approaches presented in [1, 5] use ontologies retrieved by Watson[6], and declared domain ontologies respectively. In addition, authors in [6] use WordNet senses to ground tags.

On the other hand, our multilingual sense repository was inspired by the Tagora sense repository [19]. The Tagora sense repository, similarly to Tagpedia[7], was created taking advantage of wikipedia information. However, none of these sense repositories, include multilingual information.

[4] The Spanish translation of New York
[5] `http://dbpedia.org/resource/New_York_City`
[6] `http://watson.kmi.open.ac.uk/WatsonWUI/`
[7] `http://www.tagpedia.org/`

The most related work to our proposal is [20]. In that work, authors aim at associating semantics to user tags assuming that each tag has been used in a unique sense by each user. In contrast, our work assumes that each user can use each tag in distinct senses. In addition, Tesconit *et al.* use as context the set of user tags plus those popular tags extracted from Delicious for the set of user resources. Instead, we use the co-occurring tags in the same user annotation.

We want to note that in our disambiguation activity we use the cosine similarity function instead of a custom formula used in [20], to measure the overlap of terms in the tag context with respect to the terms in each sense. The cosine function is a well studied similarity measure in Information Retrieval. Finally, this approach uses Tagpedia as sense repository. However, Tagpedia just deal with English tags, while our approach is multilingual.

3 Sem4Tags a Process for the Semantic Grounding of Cross-lingual Folksonomies

As discussed in the introduction, one of our objectives is to be able to identify the meaning of a tag by associating this tag with a resource in DBpedia. In this process, we take into account the tag context, understood as the set of user tags co-occurring when annotating a resource. It is important to note that the tag and the tags included in its context may be written in different languages. To accomplish this cross-lingual semantic grounding of tags, we propose Sem4Tags[9], a process capable of choosing the semantic entity that better defines the tag meaning in the context where it is used. Therefore, the input is a tag, its context, and, optionally, the language of the tag, while its output is the corresponding semantic entity.

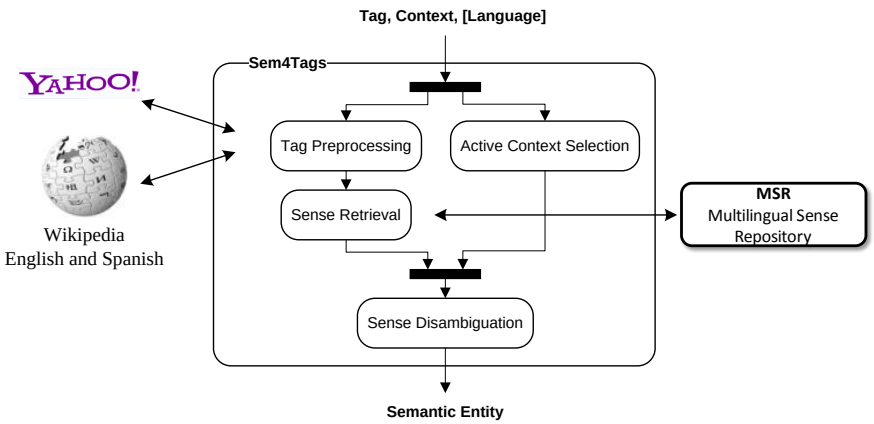

Fig. 1. Sem4Tags architecture.

Sem4Tags (see Figure 1) preprocesses the tag to find a normalized representation based on Wikipedia article titles. For doing this, we modify morphologically

the tag and use the Yahoo spelling service to find alternative representations of the tag. Simultaneously, the tags in the context are filtered to select their active context [12], that is, the set of tags in the context that are most related to the analyzed tag.

Next we query our multilingual sense repository to find the possible senses for the tag. Then in case of an ambiguous tag, a sense disambiguation activity is executed to select the sense that best represents the tag. This disambiguation activity may use the active context, as we will see in the evaluation section. Finally, the output of this process is a DBpedia resource that represents the intended meaning of the input tag. In the following section, we describe each of the activities, providing details of the technique and resources that are used, and giving some examples and intermediate results.

3.1 Tag Preprocessing

Even though tags used in folksonomies tend to converge [11], tags are written freely by users. Thus, several slightly modified tags can refer to the same concept. For instance, tags such as *nyc*, *newyork*, and *New york* might refer to New York City. Therefore, our first activity is focused on finding a tag normalized version that can lead us to better identify the main concept they refer to[8]. Our approach grounds tags to Wikipedia article titles as first step towards their normalization.

To achieve our task, first we attempt to benefit from Wikipedia redirection pages[9] in those cases where the tag has been considered as an alternative to an article title by Wikipedia users (e.g., this is the case of the *nyc* that has been considered in Wikipedia as an alternative to the *New_York_City* article). We also modify the tags to transform them to the standard notation of Wikipedia article titles: the initial letter of each word is in upper case except for in-between words such as *the, for, and* in English and *el, la, los, por* in Spanish, and words are separated using the '_' character. Finally, if after those modifications we have not found a Wikipedia article title we use the Yahoo! spelling service[10] to split concatenated words and detect misspellings. Then we transform the spelling suggestions into valid Wikipedia article titles.

In the set of tags that we handle in our evaluation, and which is described later, from the set of tags that evaluators were able to identify their meaning and language, we were capable of associate the 86.9% of tags in English, and 86.7% in Spanish to Wikipedia articles, 76.4% of the tags in English and 76.6% in Spanish required modifications to find the Wikipedia article.

3.2 Active Context Selection

As we are interested in the tag meaning in each tagging activity (i.e., the action of tagging a resource), we define the context of a tag in a resource as the set

[8] Please note that the result of this pre-processing is not necessarily the sense that we maybe looking for in all cases.

[9] `http://en.wikipedia.org/wiki/Redirects_on_wikipedia`

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

of additional tags that the user used to annotate the resource. However, many tags refer to subjective impressions of users (e.g., *my favourite*, *amazing*) or technical details (e.g., *Nikon*, *photo*) which can be useless (or even harmful) for disambiguation. Therefore, we need to select among all tags in the context those that are most related to the analyzed tag.

To carry out this selection we use a technique described in [12]. After removing repeated words and stop words from the context, we compute the semantic relatedness between each context word and the word to disambiguate. This relatedness computation is performed by using a web-based relatedness measure taking into account the co-occurrence of words on web pages, according to frequency counts, and giving a value between 0 and 1, which indicates the degree of semantic relatedness that holds between the compared words. Finally, we construct the active context set with the context words whose relatedness score above a certain threshold. We limit the cardinality of the active context to 4 (according to Kaplan's experiments [14], this is the number of words above which the context does not add more resolving power to the disambiguation).

3.3 Sense Retrieval

Our approach is inspired by the Tagora sense repository [19], which uses Wikipedia content articles as senses. We have extended this repository into MSR, a multilingual sense repository based on Wikipedia and DBpedia information (details of MSR are presented in section 3.5). Therefore, in this activity, we query MSR to find the candidate senses for a tag. In addition, from MSR we obtain for each sense the set of most frequent keywords in the corresponding Wikipedia article, along with their frequency value.

3.4 Sense Disambiguation

The sense disambiguation activity is carried out when the sense retrieval activity returns more than one sense. For instance, the *NYC* and *Nueva York* terms according to English and Spanish Wikipedias[11] have 16 and 24 possible meanings respectively. The goal of this activity is to select a sense representing the tag meaning in the context where it was used. The main idea is that the tag and its context can be compared against each one of the candidate senses, measuring the overlap of the terms in the context with the terms in the Wikipedia pages related to the senses. We carry out this process by representing the senses and the tag along with its context as vectors and then comparing those vectors using a similarity formula to find the most similar sense to the tag and its context [9].

First we create the *Vocabulary* set as the union of the top N frequent terms in each of the candidate senses. Next for each sense we create a vector in $\Re^{|Vocabulary|}$ where each position corresponds to an element in an ordered version of the Vocabulary set. The value w_i associated with the i-th position in the

[11] see `http://en.wikipedia.org/wiki/NYC_(disambiguation)` and `http://es.wikipedia.org/wiki/Nueva_York_(desambiguacion)`

vector is calculated using TF-IDF[12][22] for the corresponding i-th term in the ordered set.

Similarly, we create a vector for the tag and its context. In this case, w_i takes as value 1 if the i-th term in the ordered set appears in the tag context, and 0 if not. We compare the tag vector and each one of the sense vectors using as similarity measure the cosine function. Thus, we select the sense vector with the highest similarity value with respect to the tag vector. The retrieved information from MSR for the sense contains the DBpedia resource associated to this sense. Therefore, we return this resource as the semantic entity to associate to the tag.

3.5 Multilingual Sense Repository

MSR is a multilingual sense repository, where each sense corresponds to a Wikipedia article, and where the most frequent terms in the article are stored along with their frequency values. MSR has been built taking advantage of: 1) Wikipedia article[13] URLs as sense identifiers, and article words along with their frequency as keywords associated with the sense, 2) articles listed in disambiguation pages[14] as possible senses for ambiguous words, and 3) the explicit translations among articles[15] to link senses in languages different from English to English senses. We want to note that MSR can be easily extended to other languages due to the fact that it relies mostly on Wikipedia information, and hence its coverage will be highly related to the coverage of Wikipedia in the corresponding language.

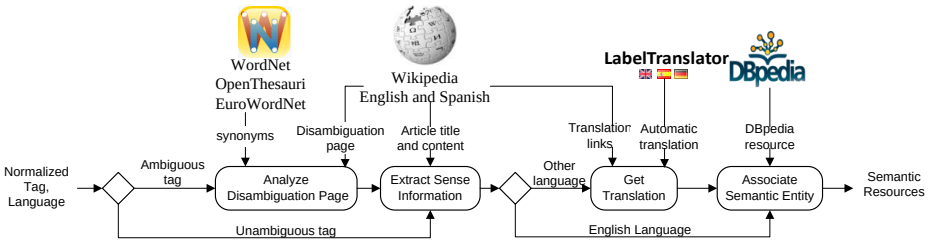

Fig. 2. Creation process of the Multilingual sense repository [9].

MSR is populated incrementally (See Figure 2). Each time a tag is presented to Sem4Tags its normalized version is queried in MSR. If there is no information in the sense repository for this tag, then the population process starts. 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 of the tag using synonyms extracted from lexical resources like EuroWordNet [21]. Otherwise, we look for a content page related to the tag. Then, for each candidate sense we

[12] TF-IDF stands for Term Frequency and Inverse Document Frequency
[13] Content pages describe subjects related to concepts, instances and named entities.
[14] Disambiguation pages list articles representing the possible uses of a word.
[15] Articles have explicit translation links into other languages.

extract the keywords and their frequency from the corresponding article. For tags in languages different than English, we look for English translations in Wikipedia and using the LabelTranslator tool[7].

Finally, we extract from DBpedia the semantic resources related to the candidate senses. In DBpedia English and Spanish Wikipedia articles are linked to DBpedia resources by means of the `page`[16] and the `wikipage-es`[17] relations. In case the `wikipage-es` relation does not exists for an Spanish Wikipedia article, we use the translation found in the previous activity and use the `page` relation. A more complete description of the creation process of MSR can be found in [9].

4 Experimental Evaluation Setup

Our evaluation has focused on determining the precision of our semantic grounding approach, considering different alternatives or decisions that can be taken in the tag processing process that we have described in section 3.

For this, we have used as test data a set of tagging activities taken from Flickr. By exploring Flickr images we found that some pictures of tourist cities were annotated with multilingual tags. Thus, we queried the Flickr API for pictures tagged with touristic places in Spain (e.g., Barcelona, Canary Island, Ibiza, etc.). We gathered a total of 764 photos uploaded to Flickr by 719 distinct users. On average those 764 photos were annotated using 12.4 tags with a standard deviation of 7.85. In addition, our data set consists of 9484 tagging activities, that is, 9484 triples of the form $\langle user, tag, photo \rangle$, where 4153 distinct tags were used. Each tag was used on average 2.28 times to annotate the pictures with a standard deviation of 5.69.

Our baseline attempts to directly match tags with Wikipedia article titles. We first preprocess tags by looking for spaces and replacing them with the '_' character. Next, for tags in English we create a URI of the form `http://en.wikipedia.org/wiki/tag`. After that, with that URI we query DBpedia for a semantic resource using the *page*[18] relation. On the other hand, with tags in Spanish we create a URI of the form `http://es.wikipedia.org/wiki/tag`. Then with that URI we query DBpedia for a semantic resource using the *wikipage-es*[19] relation.

As aforementioned, we wanted to explore the behavior of our process under different alternatives. First, we were interested in evaluating how well Sem4Tags performs when the keywords representing each sense are the most frequent terms in the whole wikipedia articles, against a more reduced set of terms extracted from article abstracts (i.e, the first paragraph describing the article content). Our hypothesis was that large Wikipedia articles can contain as frequent keywords some terms that are not necessarily related to the article main subject. In contrast, abstracts provide more concise information about the article subject and thus those terms can lead to better disambiguation results. In addition, we

[16] `http://xmlns.com/foaf/0.1/page`
[17] `http://dbpedia.org/property/wikipage-es`
[18] `http://xmlns.com/foaf/0.1/page`
[19] `http://dbpedia.org/property/wikipage-es`

wanted to evaluate if the selection of the active context leads to better disambiguation results against using the whole set of tags in the context. Thus, we evaluated the following approaches for the semantic grounding of tags:

- **Baseline**: Selection of the sense without a disambiguation activity.
- **Sem4Tags**: For each sense we use the whole Wikipedia article as source for frequent terms.
- **Sem4TagsAC**: Same as Sem4Tags including the selection of the Active Context.
- **Sem4TagsAbs**: For each sense we use the first paragraph of the Wikipedia article as source for frequente terms.
- **Sem4TagsAbsAC**: Same as Sem4Tags Abs including the selection of the Active Context.

4.1 Evaluation Campaign

We engaged 41 evaluators in the evaluation campaign. Each of them had to evaluate a set of semantic associations[20] generated by each approach, and for that evaluators were presented with the top 5 semantic entities produced by each approach[21]. We made sure that each semantic association was evaluated by at least three evaluators so that we can use those decisions taken by user majority.

For each tagging activity evaluators had to decide whether they were able to identify the semantics of the tag in that tagging activity. As context we presented to them the picture along with the other tags the user used to annotate the picture. If they were able to identify the tag meaning, we asked them to identify the language of the tagging activity[22]. In addition, users indicated if the tag correspond to a named entity (e.g., we were interested in organizations, people, places, and products). Evaluators had to evaluate the semantics associations according to their language selection. We presented to them the set of DBpedia resources (title and abstract) returned by all the approaches in the first 5 positions. They were asked to state if each DBpedia resource associated with the tagging activity was highly related (HR), related (R), or not related (N).

4.2 Metrics

To evaluate the proposed approaches, we translate the semantic association between social tags and DBpedia resources into an Information Retrieval (IR) task. Roughly speaking, the IR problem can be formulated as follows: given a (keyword based) query and a set of (text) documents, find the subset of documents that are relevant to the query. In general, a user submits a query to a search

[20] tuple of the form $\langle user, tag, photo, DBpedia\_resource, language \rangle$

[21] It is important to note that evaluators did not know where semantic associations were coming from

[22] Currently, our approach starts by considering both English and Spanish as the languages in which the tag may be available, since we do not necessarily know the language in which a tag is written

engine, and the search engine returns a ranked list of documents. In this list, not all the documents are really relevant. It is then desired that the returned list would contain as many relevant documents as possible. Moreover, since users usually focus their attention on the first results, it is also desired that relevant documents would appear in the top positions of the ranked list.

In the conducted experiment, evaluators identified which DBpedia resources were (highly) related to a given tag within the corresponding semantic context, i.e. for the annotated picture, and the presented approaches attempt to retrieve such resources. Based on the above IR problem definition, a tag (within a specific semantic context) can be considered as a query, and the set of DBpedia resources related to such tag can be interpreted as the set of relevant documents.

Aiming to empirically compare the performance of the investigated approaches, we made use of metrics widely used in the IR field. The first utilized metric is precision. In our context, for a given approach and tag, we define precision as the fraction of the DBpedia resources retrieved by the approach that are related to the tag. Since, in general, our final goal is to obtain a single related resource, we measure average precision values taking into account only the first results returned by the approaches. In the literature, this measure is called precision at one or $P@1$ [22]. For more exhaustive comparisons, we also compute $P@N$, with $N = 2, 3, 4, 5$. Furthermore, averaging the sum of $P@N$ values by the number of related resources, we define the mean average precision or MAP. The second utilized metric is recall. For a given approach and tag, we define recall as the fraction of DBpedia resources related to the tag that are successfully retrieved by the approach. Similarly to precision, we also take into consideration recall at N or $R@N$, with $N = 1, 2, 3, 4, 5$. Precision and recall are metrics that capture different aspects of the set of retrieved documents. In many situations, the use of a single measure combining precision and recall is appropriate. Thus, as our third metric, we propose to use the well known F measure, which is basically the weighted harmonic mean of precision and recall: $F = 2 \cdot precision \cdot recall/(precision + recall)$.

The previous metrics have been (and are being) used extensively by the IR community. However, they have limitations, and should be complemented by other metrics. Hence, for example, precision and recall do not take into account the usefulness of a document based on its position in the result list. To address this issue, we also compute $NDCG$ [23] and MRR [24]. $NDCG$ (Normalized Discounted Cumulative Gain) penalizes relevant DBpedia resources appearing lower in a result list. This penalization is based on a relevance reduction logarithmically proportional to the position of the relevant resources. MRR (Mean Reciprocal Rank), on the other hand, is the average of the "reciprocal ranks" for a sample of tags, where the reciprocal rank of a result list is the multiplicative inverse of the position of the first relevant DBpedia resource retrieved. To conclude, note that all explained metrics have range $[0, 1]$, and that the higher the metric value, the better performance of the retrieval approach.

5 Evaluation and Discussion

We evaluated a total of 2260 tagging activities (TAS) corresponding to 764 pictures tagged with 1112 tags[23] (see table 1). Evaluators were able to identify the semantics of 87% of the TAS (known entities). From this subset, 62.6% were considered in English and 87.7% in Spanish[24]. In addition, the number of tags considered as named entities was 53.3% in English and 55.7% in Spanish.

Table 1. Description of the dataset.

	Users	Evaluations	Evaluations/ user	Pictures	Tags	TAS	TAS/ picture
All entities	41	138063	3367.39 (±142.17)	764	1112	2260	2.96 (±0.23)
English known entities	41	30400	741.46 (±206.51)	642	659	1232	1.92 (±0.79)
Spanish known entities	41	49568	1208.98 (±152.10)	742	816	1727	2.33 (±0.74)
English known named entities	40	11872	296.80 (±152.39)	470	335	657	1.40 (±0.60)
Spanish known named entities	41	26688	650.93 (±183.01)	597	450	963	1.61 (±0.70)
English known unnamed entities	41	18528	451.90 (±120.59)	455	422	687	1.51 (±0.65)
Spanish known unnamed entities	41	22880	558.05 (±117.84)	599	533	966	1.61 (±0.70)

5.1 Precision and Recall Analysis

Tables 2 and 3 show the results obtained by the different approaches on entities marked as English and Spanish respectively. As explained in Section 3, in the experiment, each tagging activity was evaluated by three users. The users had to decide whether or not a semantic entity was relevant for a tag, and had to state the language of such tag, choosing one out four options: *English*, *Spanish*, *both* and *other*. For a given tag, based on the three users' evaluations, a semantic entity was considered relevant if at least two users stated it was *highly related* (or *related/highly related*) to the tag. There was a substantial agreement among users. Fleiss' kappa statistic [25] measuring users' agreement was $\kappa = 0.76$ (a value $\kappa = 1$ means complete agreement) for the *highly related* case, and $\kappa = 0.71$ for the *related/highly related* case. In the reported results, the former case was used because of its higher agreement level. Similar average performance results were obtained with the latter case. Precision values were higher and recall values were lower. There were more relevant entities so it was easier to accurately retrieve a relevant entity, while it was more difficult to retrieve all relevant entities. Similarly to the definition of relevance agreement, a tag (within a certain semantic context) was considered in English (Spanish) if at least two users chose *English* (*Spanish*) or *both* options. There was an almost perfect

[23] Dataset available in `http://delicias.dia.fi.upm.es/wiki/images/5/55/VW_EVALUATION.zip`

[24] Evaluators defined that the tag was written in the specific language or in both. For instance, the Madrid tag is valid both in Spanish and English.

agreement among users. Fleiss' kappa statistic was $\kappa = 0.83$. In the case of named and unnamed entities, Fleiss' kappa statistic was $\kappa = 0.85$.

Table 2. Evaluation results achieved by the different approaches for **English** entities. Wilcoxon's statistical test was conducted for MAP, P@1, R@1, F-measure, MRR and NDCG metrics. Values in underline bold (p=0.01), bold (p=0.05), and italic bold (p=0.1) indicate a statistical significance difference with values achieved by the baseline approach. Values marked with ‡(p=0.01), †(p=0.05), and *(p=0.1) indicate a statistical significance difference with values achieved by Sem4Tags approach.

	MAP	P@1	P@2	P@3	P@4	P@5	R@1	R@2	R@3	R@4	R@5	F	MRR	NDCG
All entities														
Baseline	0.78	0.88	-	-	-	-	0.78	-	-	-	-	0.28	0.88	0.81
Sem4Tags	**0.91**	0.89	0.53	0.37	0.29	0.23	0.81	0.91	0.93	0.95	0.96	**0.36**	***0.93***	**0.93**
Sem4TagsAC	**0.90**	0.90	0.52	0.36	0.28	0.23	***0.82****	0.90	0.92	0.93	0.94	**0.36**	***0.93***	**0.92**
Sem4TagsAbs	***0.84***†	0.82*	0.48	0.34	0.26	0.22	0.75*	0.85	0.89	0.90	0.92	**0.34**	0.88*	***0.87***†
Sem4TagsAbsAC	***0.86***†	0.86	0.48	0.34	0.26	0.22	0.79	0.86	0.89	0.90	0.92	**0.34**	0.90	**0.88**†
Named entities														
Baseline	0.76	0.88	-	-	-	-	0.76	-	-	-	-	0.28	0.88	0.79
Sem4Tags	**0.91**	***0.91***	0.55	0.39	0.30	0.24	***0.80***	0.91	0.93	0.94	0.96	**0.38**	***0.94***	**0.93**
Sem4TagsAC	**0.91**	***0.92***	0.55	0.38	0.30	0.24	***0.81***	0.91	0.93	0.94	0.95	**0.37**	***0.95***	**0.93**
Sem4TagsAbs	**0.86***	0.86	0.50	0.35	0.27	0.23	0.77	0.85	0.89	0.90	0.91	**0.35**	0.90*	***0.88***†
Sem4TagsAbsAC	**0.87**†	0.88	0.50	0.35	0.27	0.23	0.79	0.86	0.89	0.90	0.92	**0.35**	0.92	**0.89***
Unnamed entities														
Baseline	0.82	0.88	-	-	-	-	0.82	-	-	-	-	0.29	0.88	0.83
Sem4Tags	***0.89***	0.86	0.50	0.35	0.27	0.22	0.81	0.90	0.94	0.95	0.96	**0.35**	0.91	***0.91***
Sem4TagsAC	***0.88***	0.86	0.48	0.34	0.26	0.21	0.81	0.88	0.91	0.92	0.93	**0.34**	0.90	***0.90***
Sem4TagsAbs	0.81†	**0.77***	0.46	0.32	0.25	0.21	**0.72***	0.84	0.88	0.90	0.91	***0.33***	***0.84****	0.84†
Sem4TagsAbsAC	0.84*	***0.81***	0.47	0.32	0.25	0.21	***0.76***	0.86	0.88	0.90	0.92	***0.33***	0.87	0.87*

The results shown in the tables were obtained from those tagging activities where the associated semantic entities were known for the evaluators, and in which the corresponding tags were linked to DBpedia resources by at least one approach. The metrics explained in Section 4.2 (MAP, $P@N$, $R@N$, F, MRR and $NDCG$) were computed for each approach with English and Spanish entities. We also report the metric values for only (English/Spanish) named and unnamed entities. It is important to note that *recall* is computed assuming that the set of *all* tags relevant to a given tag is composed by the *relevant* (see definition above) entities retrieved by the investigated approaches. We cannot assure that we are able to retrieve all relevant entities but a strong representative sample of them.

Wilcoxon's statistical tests were performed to determine whether there were statistical significance differences between the metric values obtained with the baseline and the proposed approaches, and between the metric values obtained with Sem4Tags approach and its variants Sem4TagsAC, Sem4TagsAbs, and Sem4TagsAbsAC. The statistical tests were applied on those tagging activities where all approaches (including the baseline) were able to link at least one DBpedia resource. This allows us to present a more fair comparison among approaches, but implies a loss of information that hides a higher statistical evidence in the differences with metric values of approaches able to link DBpedia resources in a large number of cases.

Table 3. Evaluation results achieved by the different approaches for **Spanish** entities. Wilcoxon's statistical test was conducted for MAP, P@1, R@1, F-measure, MRR and NDCG metrics. Values in underline bold (p=0.01), bold (p=0.05), and italic bold (p=0.1) indicate a statistical significance difference with values achieved by the baseline approach. Values marked with ‡(p=0.01), †(p=0.05), and *(p=0.1) indicate a statistical significance difference with values achieved by Sem4Tags approach.

	MAP	P@1	P@2	P@3	P@4	P@5	R@1	R@2	R@3	R@4	R@5	F	MRR	NDCG
All entities														
Baseline	0.71	0.88	-	-	-	-	0.71	-	-	-	-	0.27	0.88	0.74
Sem4Tags	**0.93**	**0.93**	0.58	0.42	0.33	0.27	**0.79**	0.90	0.95	0.97	0.98	**0.41**	**0.96**	**0.95**
Sem4TagsAC	**0.93**	**0.94**	0.57	0.42	0.33	0.27	**0.80***	0.89	0.93	0.96	0.96	**0.40**	**0.96**	**0.95**
Sem4TagsAbs	**0.88**‡	0.90*	0.53	0.39	0.32	0.26	***0.76****	0.85	0.90	0.93	0.94	**0.39**	**0.93***	**0.91**‡
Sem4TagsAbsAC	**0.89**‡	0.91*	0.54	0.40	0.32	0.26	***0.77***	0.85	0.90	0.93	0.94	**0.39**	**0.94***	**0.91**‡
Named entities														
Baseline	0.67	0.90	-	-	-	-	0.67	-	-	-	-	0.27	0.90	0.72
Sem4Tags	**0.94**	**0.96**	0.64	0.48	0.38	0.31	**0.74**	0.87	0.93	0.96	0.97	**0.45**	**0.98**	**0.96**
Sem4TagsAC	**0.93**	***0.96***	0.63	0.47	0.38	0.31	**0.74**	0.87	0.92	0.96	0.96	**0.44**	**0.98**	**0.95**
Sem4TagsAbs	**0.88**‡	0.92*	0.58	0.44	0.36	0.30	***0.71****	0.82	0.88	0.92	0.93	**0.42**	**0.95***	**0.91**‡
Sem4TagsAbsAC	**0.88**‡	0.92*	0.58	0.44	0.36	0.30	***0.72***	0.82	0.88	0.93	0.93	**0.42**	**0.96***	**0.91**‡
Unnamed entities														
Baseline	0.78	0.84	-	-	-	-	0.78	-	-	-	-	0.27	0.84	0.79
Sem4Tags	**0.93**	***0.91***	0.53	0.38	0.29	0.24	***0.83***	0.92	0.96	0.98	0.98	**0.37**	**0.95**	**0.95**
Sem4TagsAC	***0.92***	***0.93***	0.52	0.37	0.28	0.23	**0.85***	0.91	0.95	0.96	0.96	**0.36**	**0.95**	**0.94**
Sem4TagsAbs	***0.89***	0.89	0.50	0.36	0.28	0.23	0.81	0.88	0.92	0.94	0.94	**0.36**	***0.92****	***0.91***
Sem4TagsAbsAC	***0.90***	0.90	0.51	0.36	0.28	0.23	***0.82***	0.89	0.92	0.94	0.94	**0.36**	***0.93***	***0.92***

Finally, note that the baseline retrieves a single semantic association for each tag. For this reason, metrics $P@N$ and $R@N$ with $N = 2, 3, 4, 5$ are not reported for that approach. Indeed, the coverage (recall) of the baseline is low in comparison to the proposed approaches, as shown in the tables. Analyzing the obtained results, the following conclusions can be drawn from our study.

- In general, the baseline obtained high performance results with tags in English and in Spanish. This fact suggests that a high percentage of the analyzed tags were used in the sense directly found by the Baseline. However, as we will discuss in Section 5.2, the baseline was able to find semantic resources for just a fraction of the analyzed data set.
- Sem4Tags and its variants perform better when dealing with Spanish tags. The amount of information in the Spanish Wikipedia compared with the English version is considerably lower[25]. Nevertheless, the Spanish version seems to contain more precise information that leads to better results.
- All approaches obtained better precision with named entities than with unnamed entities. The same observation is applicable to ranking based metrics MRR and $NDCG$. The first positions of the approach rankings tend to have more relevant results for named entities. This can be explained by the fact Wikipedia is more an encyclopedia than a dictionary, and thus named entities are a central part of the Wikipedia compared with other words.

[25] As of June 2010, the English and Spanish Wikipedia have 3,332,294 and 614263 articles respectively

- **Sem4Tags and Sem4TagsAC were the approaches that obtained the best results both in term of precision and recall**. Almost all of these results present statistical significant differences with results obtained with the baseline. Comparing Sem4Tags and Sem4TagsAC, we do not find a clear enhancement of semantic associations when exploiting the active context. In some cases, it seems that Sem4TagsAC obtains better $P@1$ and $R@1$ values, but the improvements are supported by no or low statistically evidence. This observation could be biased by the way in which statistical tests were conducted, as explained before.
- **Sem4TagsAbs and Sem4TagsAbsAC are clearly the worst approaches**. The exploitation of Wikipedia contents carried out by Sem4Tags, and Sem4TagsAC is essential to correctly associate DBpedia resources to tags. In this case, the use of the active context also seems to slightly improve precision and recall results, but again there is no statistical evidence to support this claim.

5.2 Ambiguity and Multilinguality

Using information from MSR and the Sem4Tags relevant semantic associations we can report some statistics and conclusions about ambiguity and multilinguality of tags. 61% of TAS correspond to ambiguous tags in English and 42% in Spanish[26] (see left hand side of Figure 3). In addition, for the 83% and 80% of ambiguous TAS in English and Spanish, the sense selected by Sem4Tags coincided with the Wikipedia default sense of the tag (i.e, those wikipedia pages that editors have defined to display first in case of ambiguous terms). **This shows that in the 17% and 20% of the ambiguous TAS in English and Spanish we required a disambiguation activity to select the proper sense**, what supports the need for our work.

The baseline always retrieves the default sense for a tag regardless of its ambiguity. Its high $P@1$ value can be due to the fact that in the **90% of TAS in English and 91% in Spanish the selected sense corresponds with the default sense (not ambiguous TAS plus default sense selected for ambiguous TAS)**. Nevertheless, the coverage of the baseline, defined as the number of semantic associations produced by the baseline divided by the total number of TAS, is extremely low: 27.7% in English and 19.4% in Spanish. This contrast, with the 79.1% of Sem4Tags coverage in English and 81.4% in Spanish.

Finally, tags multilinguality is clearly present in our analyzed dataset (See right hand side of Figure 3). A detailed analysis of the data shows us that 53% of pictures were tagged with tags considered valid in both languages. This fact shows that **approaches addressing tags in a unique language are not enough as to getting benefit of the whole information contained in a Folksonomy**.

[26] The average of senses was 23.3 for English and 10.35 for Spanish

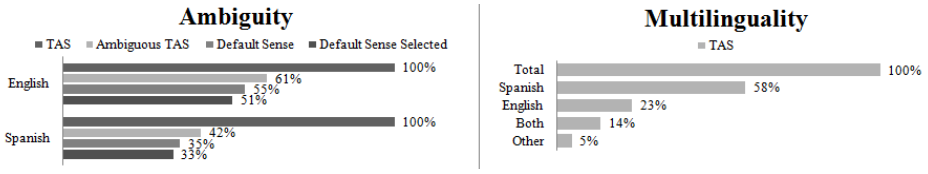

Fig. 3. Multilinguality and Ambiguity of TAS with relevant results produced by Sem4Tags

6 Future Work

The exploitation of the active context of a tag seems to improve the performance of our approach. Nonetheless, we do not obtain statistically significant evidence to support that claim. Additional evaluations focused on measuring the importance of semantic context have to be done. Analyzing the performance of our approach we find out that named entities lead to better precision. We shall carry out a more exhaustive study on this direction.

Based on the satisfactory results achieved by Sem4Tags, and the easy extension of MSR to other languages, we plan to carry out experiments with other languages so that we can analyze how distinct language characteristics affect our semantic grounding approach. In addition, though the conducted experiment includes a wide range of tags and users, they belonged to the tourist places domain. We want to run our approach using tags from different domains to compare if the achieved results hold with other kind of tags. Finally, we are interested in using our evaluation results, where users have stated that some semantic entities are related to some tags in a particular context, as an starting point to produce a testbed that allows to compare the existing approaches to identify the semantics of tags.

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. Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D.: Beyond Kappa: A Review of Interrater Agreement Measures. The Canadian Journal of Statistics / La Revue Canadienne de Statistique, (27)1, 3-23(1999)
3. Begelman G., Keller P., Smadja F.: Automated tag clustering: Improving search and exploration in the tag space. In 15th International World Wide Web Conference. Edinburgh, Scotland (2006)
4. 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)
5. Cantador, I., Szomszor, M., Alani, H., Fernández, M., Castells, P.: Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations.

In 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008). Tenerife, Spain (2008)
6. Cattuto, C., Benz, D., Hotho, A., and Stumme, G.: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In the 7th international Conference on the Semantic Web. Karlsruhe, Germany (2008)
7. Espinoza, M., Gómez-Pérez, A., Mena, E.: Enriching an ontology with multilingual information. In the 5th European Semantic Web Conference. (2008)
8. García-Silva, A., Corcho, O., Alani, H., Gómez-Pérez, A.: Review of the state of the art: Discovering and Associating Semantics to Tags in Folksonomies. To appear in The Knowledge Engineering Review (2010)
9. García-Silva, A., Corcho, O., Gracia, J.: Associating Semantics to Multilingual Tags in Folksonomies (Poster). To appear in the 17th International Conference on Knowledge Engineering and Knowledge Management EKAW. Lisbon, Portugal (2010).
10. Giannakidou, E., Koutsonikola, V., Vakali, A., Kompatsiaris, Y.: Co-Clustering Tags and Social Data Sources. In 9th International Conference On Web-Age Information Management (WAIM 2008). (2008)
11. Golder, S. A. and Huberman, B. A. 2006. Usage patterns of collaborative tagging systems. J. Inf. Sci. 32, 2 (2006), 198-208.
12. Gracia, J., Mena, E.: Multiontology semantic disambiguation in unstructured web contexts. In Workshop on Collective Knowledge Capturing and Representation (CKCaR09) at K-CAP 2009, Redondo Beach, California (USA), 2009.
13. Jäschke, R., Hotho, A., Schmitz, C., Ganter, B. Stumme, G.: Discovering shared conceptualizations in folksonomies. Journal of Web Semantics 6(1), 38–53 (2008)
14. Kaplan, A.: An experimental study of ambiguity and context. Mechanical Translation 2(9), 39–46 (1955)
15. Kim, H. L., Scerri, S., Breslin, J. G., Decker, S. & Kim, H. G.: The state of the art in tag ontologies: a semantic model for tagging and folksonomies. In the international Conference on Dublin Core and Metadata Applications. Berlin, Germany (2008)
16. Marlow, C., Naaman, M., Boyd, D., Davis, M.: HT06, tagging paper, taxonomy, Flickr, academic article, to read. In the seventeenth conference on Hypertext and hypermedia (HYPERTEXT '06). Odense, Denmark (2006)
17. Mika, P.: Ontologies are us: A unified model of social networks and semantics. Journal of Web Semantics 5(1), 5–15 (2007)
18. Qiu, Y., Frei, H.P.: Concept Based Query Expansion. In 16th ACM International Conference on Research and Development in Information Retrieval (SIGIR-93). Pittsburgh, USA (1993)
19. Tagora sense repository, `http://tagora.ecs.soton.ac.uk/tsr/index.html`
20. Tesconi, M., Ronzano, F., Marchetti, A., Minutoli, S.: Semantify del.icio.us: Automatically Turn your Tags into Senses. In Social Data on the Web, Workshop at the 7th International Semantic Web Conference. Karlsruhe, Germany (2008)
21. Vossen, P.: Eurowordnet: a Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers. (1998)
22. Baeza-Yates, R. A. Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc. (1999)
23. Jarvelin, K., Kekalainen, J.: Cumulated Gain-based Evaluation of IR Techniques. ACM Transactions on Information Systems 20(4), 422-446 (2002)
24. Voorhees, E. M.: TREC-8 Question Answering Track Report. In the 8th Text Retrieval Conference, 77-82 (1999)
25. Fleiss, J. L., Cohen, J.: The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. Educational and Psychological Measurement 33, 613-619 (1973)