Named Entity Recognition for User-Generated Tags

Nadine Ludwig
Hasso Plattner Institute for Software Systems Engineering
Potsdam, Germany
Email: nadine.ludwig@hpi.uni-potsdam.de

Harald Sack
Hasso Plattner Institute for Software Systems Engineering
Potsdam, Germany
Email: harald.sack@hpi.uni-potsdam.de

***Abstract*—Content Based Multimedia Retrieval on non-textual documents is often constrained by available metadata. User-generated tags constitute an important source of information about a resource. To enable search scenarios exceeding traditional text-based search, such as exploratory and semantic search, this textual information must be complemented with semantic entities. Due to tag ambiguities and creative neologisms automatic semantic annotation based on user tags represents a major challenge. In this work, we show how to adopt context information and ontological knowledge to automatically assign semantic entities to user-generated tags for video data. Thus, a sophisticated semantic search on semantic entities is enabled. The algorithm combines co-occurence and link graph analysis using Linked Data. Also, a definition of context reliability in audio-visual content is described.**

***Keywords*-named entity recognition; disambiguation; user-generated tags**

I. Introduction

Social bookmarking services, such as Delicious[1], constitute their success from the various community functionalities, first and foremost the ability of users to tag their own and other resources. In this way, a huge amount of valuable user-generated metadata is created. This metadata is essential to enable an efficient search within a portal, especially if the resources are non-textual resources, such as videos or images.

However, textual tags can only be used for a full text search on the tagged resources, because they don't provide explicit semantics. To enable innovative search scenarios, such as semantic and explorative search [1] user-generated metadata has to be annotated with additional semantic metadata.

User-generated tags together with other metadata form a special semantic context. On the one hand, tags offer a higher reliability than automatically extracted metadata, as e. g., text from Optical Character Recognition (OCR) or Automatic Speech Recognition (ASR). On the other hand, tags are unstructured information and do not hold supplementary information, which can be used for a disambiguation process, such as structured metadata (e. g., "Speaker: Albert Einstein"). Also, tags cannot be dealt with in the same way as running descriptive text.

[1]http://www.delicious.com

This paper addresses the characteristics of user-generated tags and introduces a novel approach to enhance text-based tags with context-based semantic annotations to enable a sophisticated explorative search on semantic entities demonstrated on the example of audio-visual content. The approach considers the distinguishing characteristics of tags to continuous text. Also, we depict a context definition for audio-visual content and draw conclusions on the informative value of tags in different context levels. The described algorithm uses Linked Data[2] - in particular DBpedia [2] entities - to detect semantic relationships between entity candidates. This semantic analysis is combined with a co-occurence analysis on Wikipedia[3] text corpora. Thus, a novel technology to recognize and disambiguate semantic entities is presented.

The rest of the paper is structured as follows: Sect. II summarizes related work on tag characteristic identification and Named Entity Recognition (NER). In Sect. III the overall approach on assigning semantic entities to user-generated tags is described in detail and in Sect. IV the evaluation of the approach is discussed. Sect. V concludes the paper with an outlook on future work.

II. Related Work

A. Tags & Tag Characteristics

User-generated tags can be characterized as keywords, category names, or metadata. Every user of a portal that offers tagging functions can tag any resource within the limit of user-specified permissions [3]. Users don't follow any formal guidelines, which results in a large variety of how they tag resources. Resources can be tagged with any term that - from the user's point of view - represents a relationship between the resource and a concept [4]. Within three identitified categories of intended audience of tags (Self, Family & friends, Public), the category Public is ranked as most important motivation for tagging [5].

Golder et al. identified seven different tag functions [6]. Most of these functions imply that tags can describe a resource on different levels of abstraction, e. g., tags can explicitly name an entity as well as be descriptive regarding the category the tagged resource belongs to. In this way a

[2]http://linkeddata.org/
[3]http://wikipedia.org

semantic search both on entity and category level can be enabled by semantically enriched user-generated tags.

B. Named Entity Recognition

The most challenging problem on mapping user-generated data to semantic entities is the existence of ambiguous names. Ambiguity results in a set of entity candidates, which have to be interpreted to identify the appropriate candidate for the given context. Related work fields are amongst others word-sense disambiguation in text documents, named entity (reference) resolution, and feature based entity matching [7]. The presence of assumed same named entities in different data sets of the Linked Open Data Cloud (LOD) [8] necessitates similarity based comparison of those entities and their associated properties [9]. In the context of named entity resolution in text documents semantic information needed for disambiguation of entity candidates has to be extracted automatically and compared to adequate knowledge resources [10]. Further research approaches are using the LOD cloud as RDF graph to find relations between entities co-occurring in a text. This is supported by the hypothesis that disambiguation of co-occurring elements in a text can be obtained by finding connected elements in an RDF graph [11].

III. Method

According to a study about structure and characteristics of folksonomy tags [12] an average of 83% of user-generated tags are single terms. Also, an average of 82% of the reviewed tags are nouns. Based on these results, we ignore tag practices for composite terms, such as camel case (”barackObama”) and consider tags as subjects or categories describing a resource. As a tag may also be part of a group of nouns representing a single entity (”flying machine”,”albert einstein”) the tags stored as single words without any given order have to be combined in term groups of two or more terms to enable a mapping to all appropriate entities. Hence, each simple tag or group of tags within a given context may represent a distinct entity. The term combination process and subsequent mapping of terms and term groups to entities are described in Sect. III-B.

To disambiguate ambiguous terms we combine two methods: a co-occurence analysis of the terms in the context of Wikipedia[4] articles and an analysis of the page link graph of the Wikipedia articles of entity candidates. The scores for both analysis steps are calculated to a total score.

A. Context Definition

Metadata exists in a certain context and has to be interpreted according to this context. For tags of audio-visual content we identified three dimensions:

- temporal dimension,
- user-centered dimension, and
- spatial dimension.

In the temporal dimension a context can be defined as the entire video, a segment or a single timestamp in the video. The user-centered dimension classifies a context by how many users have created the metadata - only tags by a distinct user or all tags regardless of any user. The spatial dimension defines a context by where in a frame tags occur. Thus, tags in the same region of a video frame are considered as related to each other. In the current approach we did not consider this context dimension, because our test setting does not hold any spatial information. Fig. 1 shows the combination of the three dimensions of context for metadata in audio-visual content and the interpretation regarding the informative value of a context.

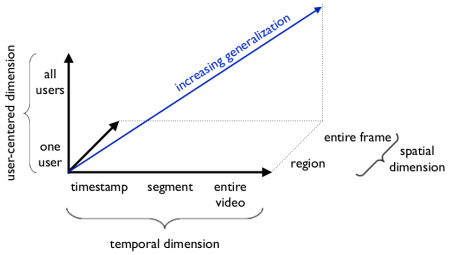

Figure 1. Dimensions of context definition in audio-visual content

To describe our approach we use a sample context of our test set (c.f. Sect. IV). This sample context is composed of tags by only one user at a given timestamp in the video. This sample context is chosen from a video[5] of a presentation by Dr. Garik Israelian at the TED conference[6]. Our sample context consists of the tags ”hubble”, ”spitzer”, ”carbon”, ”dioxide”, ”methan”, ”co2”, and ”water”.

B. Preprocessing

Term Combination: Our combination algorithm considers all tags of a specified temporal context and generates every possible combination of at most three terms within the context in any order. Thus, we make sure to combine groups of single terms that belong together. The number c of possible combinations is calculated as follows:

$$c = \sum_{k=1}^{j} \frac{n!}{(n-k)!}$$

About 90% of the DBpedia labels consist of at most three words, but less than 5% consist of 4 words. Due to these numbers and performance issues we have decided to limit the number of terms to be combined to three. For our sample context containing 7 tags and at most 3 terms in

[4]we use Wikipedia instead of other text corpora, because every DBpedia entity candidate can be easily referred to an Wikipedia article as associated co-occurence text base

[5]http://yovisto.com/play/14415

[6]http://www.ted.com

a combination ($j = 3$), 259 combinations are generated. Subsequently, in this paper by terms we will refer to single terms as well as to valid term groups.

Term Mapping: The terms then are mapped to distinct semantic entities. For our approach we use entities of the DBpedia. DBpedia provides labels for the identification of distinct entities in 92 languages. We use English and German as well as Finnish labels, as we have noticed that neither English nor the German labels contain important acronyms as labels, but the Finnish language version does. As tagging users prefer to keep it short and simple [4], resources dealing with "Domain Name System" would rather be tagged with "DNS" than "Domain Name System".

After simple string matching of the terms of the context to DBpedia entities, the URIs are revised for redirects and disambiguation URIs. That is, concerning URIs are replaced by their redirects resp. the URIs they link to as disambiguation URIs. For our sample context overall 120 candidates are mapped to 8 terms. These entity candidates have to be disambiguated within the given context. This disambiguation process is described in the next sections.

C. Co-occurence Analysis of Context Terms in Wikipedia Articles

To find the appropriate entity for a term of the context the disambiguation is processed for every entity candidate mapped to the term. In the first step, we use the Wikipedia article referring to the entity candidate to count occurence of all the other terms in the context of the term currently processed (subsequently, this analysis step is referred to as CA). The score for an entity candidate is calculated as follows:

$$C(t) = \{t_j\}, j = 1...k$$
$$W(uri(t)_i) = \{w_r\}, r = 1...|W(uri(t)_i)|$$

t is the term currently disambiguated. $C(t)$ is the set of terms in the context in which t has to be disambiguated. $W(uri(t)_i)$ is the set of all terms in the Wikipedia article for the current entity candidate $uri(t)_i$ of the term t. To calculate the CA score the number ($counter_{cooc_i}$) of how often all other terms of the context occur in the article for the entity candidate is determined as:

$$counter_{cooc_i} = \sum_{j=1}^{k} \sum_{r=1}^{|W(uri(t)_i)|} \delta(t_j, w_r)$$

with $\delta(x, y) = \{ {}^{1:\ x=y}_{0:\ else}$.

Finally, the CA score is calculated as follows:

$$score_{CA_i} = counter_{cooc_i} \cdot \frac{|W(uri(t)_i) \cap C(t)|}{|C(t)|}$$

D. Link Graph Analysis of Relationships between Entities

We assume entities that are related to each other are also linked by means of their Wikipedia articles. Thus, for this analysis step we evaluate the link graphs for the entity candidates of a context. Subsequently, this analysis step is referred to as WA.

For our approach we have identified three different link types that describe certain relationships between entities. The link types are shown in Fig. 2 in descendent order for their strength of relationship between the relevant entities.

Link types b) and c) are links with a path length of $w = 2$. That means, these entities are linked through a node, which also is an entity. E. g., Albert Einstein and Gottfried Leibniz both have incoming and outgoing links to the Berlin Academy of Sciences, but they are not directly linked in their Wikipedia articles. So, these two entities are linked with a link type b).

There are some entities in Wikipedia, that refer to numerous other entities and that are referred to by lots of other entities. We ignored these entities with the highest in- and outdegrees (such as "United States"[7] with over 300.000 incoming and almost 1.000 outgoing links), because entities that are only linked through such a highly frequented hub are probably not closely related to each other.

The WA detects connections between the entity currently processed and the entity candidates of the other terms in the context. A score for every link type is calculated similar to the calculation of the score in the CA.

We count the entity candidates the processed candidate is linked to. For link types b) and c) we also count the number of different paths between two candidates. We calculate the score for direct links as follows:

$$counter_{dlinks_i} = \sum_{j=1}^{k} \sum_{l=1}^{m} |uri(t)_i \rightarrow uri(t_j)_m|$$
$$score_{dlinks_i} = \frac{|t \rightarrow t_k|}{|C(t)|} \cdot counter_{dlinks_i}$$

$counter_{dlinks_i}$ is the number of candidates the processed candidate ($uri(t)_i$) is linked to directly.

With this calculation we achieve to get higher scores for entity candidates that are linked to only one of the candidates of the other terms. Such candidates have fewer links, but these links are more explicit. An entity candidate, that is linked to more than one of the candidates of a specific term in the context is much less relevant, because these links might reveal ambiguity again. The ranking we achieve by our score calculation is shown in Fig. 3. "uri 1" is linked to one entity candidate of every term in the context. That implies, that this entity candidate is strongly related within this context. Also, relationships of this candidate to the other terms in the context are not ambiguous as the candidate is

[7] http://dbpedia.org/resource/United_States

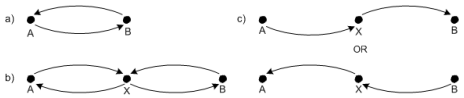

Figure 2. Three different types of Wikilinks: a) direct links, b) symmetric links through same node (symlinksl2), c) links through a node, but not symmetric (simplelinksl2)

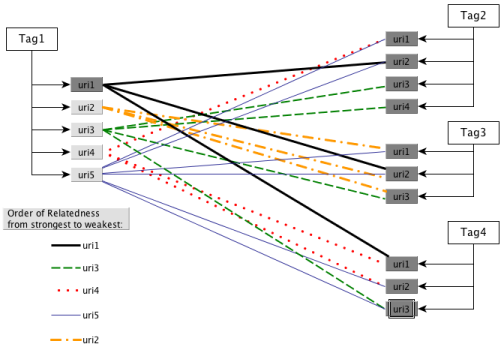

Figure 3. Ranking of relationship of entity candidates via Wikipedia page links

only linked to one of the candidates of each term. ”uri 5” is also strongly related within this context, but the links are ambiguous as the candidate is linked to two candidates of each term. ”uri 2” has the same number of links to other entity candidates as ”uri 1” in this context, but all links refer to the same term. This candidate is the least related candidate and the links are ambiguous as the candidate is linked to three different entity candidates of the same term.

The scores for link types b) and c) are calculated as follows:

$$score_{sw2_i} = \frac{|t \to t_k| \cdot p}{|C(t)| \cdot counter_{sw2}}$$

$$score_{lw2_i} = \frac{|t \to t_k| \cdot p}{|C(t)| \cdot counter_{lw2}}$$

$counter_{sw2}$ and $counter_{lw2}$ are identical to the calculation for direct links. p is the number of different paths between two linked entities.

Every (normalized) link score is weighted for a total link score according to the link type ranking. To determine these weights we have made a set of test runs and identified the best results using the weights as follows:

$$score_{WA_i} = 0.45 \cdot score_{dlinks_i} + 0.30 \cdot score_{sw2_i} + 0.25 \cdot score_{lw2_i}$$

As direct links remark the strongest connection of these three link types the direct link scores ($score_{dlinks_i}$) are weighted highest. Symmetric links with a path length $w = 2$ ($score_{sw2_i}$) are weighted a bit higher as links between two entities through a node just in one direction ($score_{lw2_i}$).

E. Evaluation of Scores

The scores of the WA and the CA for every entity candidate are weighted and added to a total score for the combination of both analysis steps. Similar to finding the weights for the different link types we have made a set of test runs to identify the most appropriate weights for the total score.

First, we have evaluated the analysis steps separately and noticed, that the CA has been slightly worse in recall and precision than the WA (CA: Recall=69% Precision=14%; WA: Recall=78% Precision=16%), but the results contained entities that were not in the result set of the WA. Thus, we have combined both analysis steps and have varied the weights for the scores for both steps successively and have identified a weight of 40% for the CA score and 60% for the WA score as best values for the combination of these analysis steps in our approach. The test sets we have used for the test runs are described in the following section.

IV. Evaluation

Along with the propellent development of the semantic web, NER and a subsequent automatic semantic annotation of textual documents are essential research areas. Recently, many approaches have been developed and provided as online service or API, as e. g., Beycoo[8], LingPipe[9], Bueda[10], or DBpedia Spotlight[11]. Unfortunately, most of these services use proprietary namespaces for their entities, which prevents a direct comparison of the approaches. Also, no test data set as benchmark for our purpose exists. For this reason, we have composed two test data sets of tags from the Yovisto[12] video search engine.The ground truth for these test sets has been created manually by colleagues and students of our research group[13].

First, we have evaluated our algorithm against our ground truth. We have achieved the highest recall with a threshold of 10% of the highest score for all entity candidates of a term. The precision is accordingly low for this evaluation. By only assigning the entity candidates with the highest score for a term we have achieved an F_1-measure of 69% for test set 1 resp. 54% for test set 2. The detailed results for this evaluation are shown in Table I and II. The lower precision results from the fact that our approach assigns entities to term combinations that have not necessarily been annotated for the relevant context in the ground truth, e. g., entities have been assigned to ”carbon”, ”dioxide”, as well as ”carbon dioxide”.

[8]http://www.beycoo.com/demo

[9]http://alias-i.com/lingpipe/web/demo-ne.html

[10]http://www.bueda.com - API phased out as product in 02/2011 and is no longer available online

[11]http://dbpedia.org/spotlight

[12]http://www.yovisto.com

[13]The data set and the ground truth is described in detail and can be downloaded on http://yovisto.com/labs/ner/

As DBpedia Spotlight uses the same namespace as our algorithm, we have used the Spotlight API to accomplish a NER on our test sets and evaluate our algorithm. Since DBpedia Spotlight requires running text as input, we have made sure our sample contexts are not containing single terms that belong together but were tagged in the wrong order. This would have given our approach an advantage, because we do consider all possible term combinations as described in Sect. III-B. For the comparison of our approach with DBpedia Spotlight we have only assigned entity candidates with the highest score to a term. As shown in Table II our approach has scored a significantly higher recall and also a higher precision for our test sets.

Table I
EVALUATION RESULTS FOR 2 TEST SETS OF TAGS - ONLY HIGHEST SCORE

	50 Segments (256 Tags)	50 Timestamps (315 Tags)
Original Mappings	11794 entity candiates (9-1224 candidates per context)	7562 entity candidates (13 - 1282 candidates per context)
Assignments	300 Entities	485 Entities

Table II
COMPARISON OF OUR NER APPROACH AND DBPEDIA SPOTLIGHT FOR FIRST AND (SECOND) TEST SET

	Spotlight	HPI
Recall	39% (42%)	78% (81%)
Precision	34% (39%)	64% (41%)
F_1-measure	36% (40%)	69% (54%)

V. CONCLUSION & FUTURE WORK

We have introduced an approach to annotate online resources semantically by using user-generated tags and mapping them to semantic entities. This approach is able to determine relationships of entity candidates for the tags in a given context. These relationships are based on simple statistical measures, such as occurence of the context tags as well as on semantic relationships derived by link graph analysis.

Ongoing research is focussed on the improvement of precision. Compared to DBpedia Spotlight our results are significantly better in both recall and precision.

In the first place, future work will address the analysis of additional metadata useful for the disambiguation of tags. The context can be extended by adding static metadata assignments. Thus, the disambiguation process is enhanced in terms of reliability. In this way, a NER workflow based on automatically assigned textual metadata can be processed successively from reliable to less reliable metadata.

Furthermore, a fine-granular method to combine tags of a context should avoid assigning entities to resources that are not meaningful for the given context and higher the precision. Also, in this approach we did not consider tag frequency and its relations to different users. Terms tagged by many users for a resource should be scored higher than terms only tagged by fews users, because these terms seem to be more relevant for the resource than others. In this way a tag ranking can be calculated to be used for the disambiguation process. NER is essential to enable a semantic search and its quality has a direct impact on the quality of a semantic search.

REFERENCES

[1] Waitelonis, J., Sack, H.: Towards exploratory video search using linked data. Multimedia Tools and Applications (2011) 1–28 10.1007/s11042-011-0733-1.

[2] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web. Volume 4825 of LNCS. Springer Berlin / Heidelberg (2007) 722–735

[3] Sawant, N., Li, J., Wang, J.Z.: Automatic image semantic interpretation using social action and tagging data. Multimedia Tools Appl. **51**(1) (2011) 213–246

[4] Tonkin, E., Guy, M.: Folksonomies: Tidying up tags? D-Lib **12**(1) (January 2006)

[5] Nov, O., Ye, C.: Why do people tag?: motivations for photo tagging. Commun. ACM **53** (July 2010) 128–131

[6] Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. J. Inf. Sci. **32** (April 2006) 198–208

[7] Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. **41**(2) (2009)

[8] Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked data on the web. In: Proc. of the 17th Int. Conf. on World Wide Web, ACM (2008) 1265–1266

[9] Stoermer, H., Rassadko, N.: Results of OKKAM feature based entity matching algorithm for instance matching contest of OAEI 2009. In Shvaiko, P., Euzenat, J., Giunchiglia, F., Stuckenschmidt, H., Noy, N.F., Rosenthal, A., eds.: OM. Volume 551 of CEUR Workshop Proceedings. (2009)

[10] Pilz, A., Paaß, G.: Named Entity Resolution Using Automatically Extracted Semantic Information. In: Workshop on Knowledge Discovery, Data Mining, and Machine Learning. (2009) 84–91

[11] Kleb, J., Abecker, A.: Entity reference resolution via spreading activation on RDF-graphs. The Semantic Web: Research and Applications (2010) 152–166

[12] Spiteri, L.F.: The use of collaborative tagging in public library catalogues. Proceedings of the American Society for Information Science and Technology **43**(1) (2006) 1–5