A Hybrid Approach for Spotting, Disambiguating and Annotating Places in User-Generated Text

Karen Stepanyan, George Gkotsis, Vangelis Banos*, Alexandra I. Cristea and Mike Joy
Computer Science Department, University of Warwick, Coventry, United Kingdom
*Department of Informatics, Aristotle University of Thessaloniki, Greece
K.Stepanyan@warwick.ac.uk, G.Gkotsis@warwick.ac.uk, *Vbanos@gmail.com,
A.I.Cristea@warwick.ac.uk and M.S.Joy@warwick.ac.uk

ABSTRACT
We introduce a geolocation-aware semantic annotation model that extends the existing solutions for spotting and disambiguation of places within user-generated texts. The implemented prototype processes the text of weblog posts and annotates the places and toponyms. It outperforms existing solutions by taking into consideration the embedded geolocation data. The evaluation of the model is based on a set of randomly selected 3,165 geolocation embedded weblog posts, obtained from 1,775 web feeds. The results demonstrate a high degree of accuracy in annotation (87.7%) and a considerable gain (27.8%) in identifying additional entities, and therefore support the adoption of the model for supplementing the existing solutions.

Categories and Subject Descriptors
H.2.8 [**Information Systems**]: Database Applications;
I.7 [**Computing Methodologies**]: Document and Text Processing

Keywords
Semantic Annotation; Named Entity Recognition; Geolocation

1. INTRODUCTION
One of the fundamental characteristics of the Semantic Web is the presence of machine-understandable annotations. Automated semantic annotation of web resources is necessary to accelerate the advancement of the Semantic Web. Enriching content with links to DBpedia resources is one of the common approaches for annotating semantic entities in the textual Web content.

Despite the considerable progress in the area of natural language processing, the performance of the available annotation tools often fails when used with (primarily) colloquial user-generated content [1]. This means that the solutions that were developed and tested with formal texts and news corpora are less effective when applied to informal or laconic writing styles that are common to weblogs and other social media sources. Given the remarkable proliferation of user-generated content, automated annotation of social media deserves a greater attention. We argue that the increasing availability of *geolocation data, embedded in social media, can be used to improve the automated process of semantic annotation of texts*. The main contribution of this paper is a new hybrid approach that takes into account embedded geolocation data to supplement the current state-of-the-art approaches for annotating user-generated content.

Previous research on the subject of semantic annotation has resulted in the development of a range of commercial and open source tools such as Stanford NER, OpenCalais, AIDA and DBpedia Spotlight. These tools employ various techniques to determine the meaning of and annotate textual fragments. However, they solely rely on processing textual content and disregard embedded metadata such as geolocation. Hereby we propose a hybrid solution and demonstrate the benefits of integrating geolocation for spotting and disambiguating places in user-generated content.

Traditionally, the solutions for semantic annotation were bound to developing a catalogue of entities, for instance from WordNet or Wikipedia, and used the associations between the entities for disambiguation. Prior work in this area used Wikipedia content, links and intra-wiki links (anchor text), redirect pages and category hierarchies [2, 3] for distinguishing between potential entities. Windowing strategy, where only a certain number of nearby words are taken into account, constitute a major limitation that lead to a significant loss of discriminative features and problems. More recent developments include the use of topic models [4] and proximity [5] that enable better performance compared to earlier approaches. However, the use of metadata, such as geolocation, can further improve the solutions for annotating places and toponyms. We describe and evaluate the proposed approach below.

2. GEOLOCATION-AWARE SEMANTIC ANNOTATION MODEL
We name the proposed model *GeoAnno*. The model consists of the following steps. It (a) takes HTML documents, which contain embedded geolocation data as input; (b) identifies associations between the published post and nearby places (using embedded geolocation and a knowledge base such as Google Places or GeoNames); (c) exploits each found association to spot relevant text faces and disambiguate entities; and (d) returns the annotated content as XML. We present a hypothetical example to illustrate this (see Figure 1).

Figure 1: Example of the annotation process and notation.

Given a weblog post with geolocation it is possible to retrieve a list of nearby *places* (e.g. $p_1, p_2 \ldots p_i$). To identify whether nearby places are associated with the textual content of the weblog post we calculate the similarity of each place name with the fragments of content and refer to the measure of similarity as *association score* (e.g. $a_1, a_2 \ldots a_i$). Once the association scores are calculated, we identify the *text faces* for each of the associated *place* (e.g. f_1p_1, f_2p_1 and f_1p_2). The text faces for the illustrated

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

example will be: f_1p_1 = 'Cafe Brazil'; f_2p_1 = 'Brazil's'; and f_1p_2 = 'Chelsea'. Finally, once the text faces have been identified, they are linked with *references to resources* (e.g. r_1 and r_2) (if any) from a knowledge base such as DBpedia.

2.1 Identifying Associated Places

We use Google Places due to an easy-to-use API and available developer libraries. After retrieving a list of place names, we identify whether the content of the weblog post is related to the nearby places. Calculation of the scores for text matching is adapted from cosine similarity. The use of positive scores as a threshold for two-grams performs robustly well. This process is described in Algorithm 1.

Algorithm 1: Identification of Association.

```
   /* Retrieve posts for annotation*/
1: List <Post> posts = CALL getPosts()//All posts with geolocation
   /*Retrieve nearby locations for each post and  calculate association scores */
2: FOR each post in posts
   /*Retrieve nearby places as a List<Place> places*/
3: CALL getNearbyPlaces(geoPoint) returning List<Place>
   places
      /*Calculate association score for each place*/
4:    FOR each place in places
5:     CALL calculateAssociationScore(post, place)
        returning float score
       /*Index post, place and score*/
6:     CALL index(post, place, score)
7:    END FOR
8: END FOR
```

2.2 Spotting Relevant Text Faces

Once the association between the post and specific nearby places is established, we identify the fragment of the text from the post, referred here as text face, to use it for annotation purposes. Rather than relying on NER approaches for identifying places that are mentioned in the text we employ techniques for text matching between the earlier discovered associated places and specific fragments of text. This approach is justified given the limitation of NER tools for processing colloquial texts.

2.3 Disambiguation and Linking

After spotting the relevant text faces it is necessary to disambiguate these when more than one potential candidate for annotation is found. This task is often necessary when annotating places that share a common name, for instance 'St. Paul's Cathedral', but refer to different entities. The use of geolocation can supplement the current approaches and aid distinguishing the St. Paul's Cathedral in London (UK) from that of in Melbourne (Australia).

3. EVALUATION

The evaluation of the proposed model was conducted using 1,775 web feeds with geolocation points embedded in 3,165 weblog posts. We evaluated the results of the semantic annotation using the CrowdFlower.com crowd sourcing system, following the methodology employed by Milne and Witten [2] for evaluating the performance of an annotation system. The workers were allocated a set of annotated weblog posts (311), which contained a single place reference at a time (334 places in total). The task of the workers was to identify whether the linked article corresponded to the highlighted text. A set of 31 questions was marked as golden to identify untrusted workers, and 136 weblog posts were incorrectly annotated with links to *nearby* but unrelated places to eliminate possible bias. The results demonstrate (see Table 1) that the accuracy of annotation was as high as 87.7%, which is comparable to the average performance (81.3%) of DBPedia Spotlight for geopolitical entities (see [6]).

Table 1: Crowd Flower evaluation results.

Judgment Result	*Judgments (%)*	*Avg. Confidence*
Correct	87.7%	0.97
Incorrect	11.4%	0.79
Ambiguous	0.9%	0.45

We compared the performance of DBpedia Spotlight with the developed prototype. The proposed model identified 209 additional resources that were not captured by DBpedia Spotlight (see Table 2). The additional number of distinct resources corresponds to a 27.8% gain in capturing additional annotations using the proposed hybrid approach. Furthermore, the comparison of sampled annotated resources demonstrates a higher granularity in annotation results. For example, our model annotates *Penang Hill* as a *resort* in *Penang district*, rather than annotating *Penang district* itself.

Table 2: Overlap and difference between the resources identified by DBpedia Spotlight and the proposed model.

Unique Annotations	*Number of Annotations for 311 posts*	
	DBpedia Spotlight	*GeoAnno (proposed)*
Distinct:	751	209
Shared:	86	86
Total:	837	295

4. CONCLUSIONS

In this paper we presented a model for geolocation-aware semantic annotation of social media content and evaluated its performance. The evaluation demonstrated a considerable gain (27.8%) in improving the performance of semantic annotation of places. It performed with a high accuracy (87.7%) and offered a finer granularity. Hence, we suggest integration of the model into the existing tools for improved performance. The application of the model can enable automated annotation of user-generated content that is *already* published or is *being prepared* for publication.

5. ACKNOWLEDGMENTS

This work was conducted as part of the BlogForever project co-funded by the European Commission FP7, grant agreement No.269963. The authors of this paper would like to thank Russell Boyatt (DCS, University of Warwick) for his continued technical support and feedback.

6. REFERENCES

[1] Ritter, A., Clark, S., Mausam and Etzioni, O. Named entity recognition in tweets: an experimental study. In *the Empirical Methods in NLP*. Edinburgh, UK. ACM, 2011.

[2] Milne, D. and Witten, I. H. Learning to link with wikipedia. In *the Information and knowledge management.* Napa Valley, California, USA. ACM, 2008.

[3] Kulkarni, S., Singh, A., Ramakrishnan, G. and Chakrabarti, S. Collective annotation of Wikipedia entities in web text. In *the SIGKDD international conference on Knowledge discovery and data mining.* Paris, France. ACM, 2009.

[4] Kataria, S. S., Kumar, K. S., Rastogi, R. R., Sen, P. and Sengamedu, S. H. Entity disambiguation with hierarchical topic models. In *the SIGKDD international conference on Knowledge discovery and data mining.* San Diego, USA. ACM, 2011.

[5] Sen, P. Collective context-aware topic models for entity disambiguation. In *the World Wide Web*. Lyon, France. ACM, 2012.

[6] Mendes, P. N., Daiber, J., Jakob, M. and Bizer, C. Evaluating dbpedia spotlight for the tac-kbp entity linking task. In *the Text Analysis Conference*. Gaithersburg, Maryland, USA. 2011.