SINAI at Placing Task of MediaEval 2010 *

José M. Perea-Ortega, Miguel Á. García-Cumbreras, L. Alfonso Ureña-López and Manuel García-Vega
SINAI Research Group, Computer Science Department
University of Jaén
23071 - Jaén, Spain
{jmperea,magc,laurena,mgarcia}@ujaen.es

ABSTRACT

In this paper we present a basic approach for assigning geographical coordinates to videos uploaded to Flickr[1]. Our approach combines geographical entities recognition based on the annotations provided for each video and information retrieval on textual tags of Flickr images provided for development purposes. However, we have not used the low-level visual features extracted from keyframes and training images. The poor results obtained show that our geographical entity recognizer must be improved to solve the spatial ambiguity present in most of the user annotations of each video. In addition, some problems have been identified during the processing of text labels from videos, such as words in capital letters that are not proper nouns or geographical entities or the use of geographical entities misspelled.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

Keywords

Video localization, Geographical coordinates, Geotagging, Flickr

1. MOTIVATION

The availability of Internet sites that contain a large number of videos is constantly growing. Most of these videos have been taken somewhere on earth and increasingly they are annotated with various forms of information including a wide variety of textual tags such us description, geographical location, time, name of the video owner, keywords, etc. In this paper we address the challenge of geotagging videos, uploaded on Flickr, using the textual annotations of each video. Our work can be considered a starting point for geotagging videos, applying a basic approach based on the use of Geo-NER [6] on the textual annotations. Geo-NER is a geographical entity recognizer that makes use of two external geographical knowledge resources, Wikipedia[2] and GeoNames[3].

2. RELATED WORK

Geotagging is the process of adding geographical identification metadata to various media such as photos, videos, websites or RSS feeds. The related work of geotagging photos is useful for geotagging videos, since, in most cases, both use the textual annotations to determine their geospatial information. The literature related to geotagging of photos is more extensive than those related to the videos. Serdyukov et al [7] investigate generic methods for placing Flickr photos using the textual annotations provided by the users. They propose a language model based entirely on these annotations. Crandall et al [3] present a system to place images on a map with a combination of textual and visual features, using a corpus of 20 million images crawled from Flickr. Hayes and Efros [4] propose visual features from Flickr images to predict the geographical information using a nearest-neighbour classification method. The remaining related literature can be divided into three main areas: spatial data mining of user-generated content [1], finding the geographical focus of web pages [2] and toponym resolution [5].

3. DESCRIPTION OF THE TASK

The *Placing* task of MultimediaEval 2010[4] requires participants to assign geographical coordinates (latitude and longitude) to each of provided test videos uploaded on Flickr. The user is free to use any kind of metadata of the videos and any external resources. In addition, organizers provide resources such as a set of Flickr images, video frames with metadata, visual features or geotags. In MultimediaEval 2010, 5125 videos in MP4 format were released for development purposes and 5091 for test. All videos included lots of metadata in an associated XML file with self-descriptive tags like *title*, *description*, *keywords*, *location*, etc. The evaluation of the results was done by calculating the distance from the point assigned by the Flickr user to the predicted point assigned by the participant and counting how many videos were correctly placed within various threshold distances: 1, 5, 10, 50 and 100 kilometers.

*This work has been partially supported by a grant from the Spanish Government, project TEXT-COOL 2.0 (TIN2009-13391-C04-02), a grant from the Andalusian Government, project GeOasis (P08-TIC-41999), and two grants from the University of Jaen, project RFC/PP2008/UJA-08-16-14 and project UJA2009/12/14.

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

[2] http://www.wikipedia.org/

[3] http://www.geonames.org/

[4] http://www.multimediaeval.org/placing/placing.html

MediaEval 2010 Workshop, October 24, 2010, Pisa, Italy

Run name	Geo-NER strategy	Collection	1 km	5 km	10 km	50 km	100 km
exp1maxpop	*max-pop*	*photos-meta*	**187**	**261**	292	771	1000
exp2maxpop	*max-pop*	*XML*	14	58	81	565	792
exp3firstl	*first-loc*	*photos-meta*	183	257	**295**	**858**	**1085**
exp4firstl	*first-loc*	*XML*	14	59	89	656	881

Table 1: Number of videos correctly predicted by SINAI in Placing task

4. EVALUATION OF RESULTS

Our approach is based on two main modules: firstly, the geographical entities from *title* and *description* tags are detected by Geo-NER [6]. Secondly, if no entity has been recognized, the system retrieves geospatial information from metadata provided for development purposes. The *title* and *description* tags were used as a query and Lucene[5] was used as a search engine. Two collections were built for indexing: the *XML* collection, generated from the 5125 XML files provided for the organizers, and the *photos-meta* collection, generated from the *tags* and geospatial information of the Flickr images provided for development purposes. When more than one geographical entity is detected by Geo-NER, we apply two main strategies: to select the entity with more population (*max-pop* strategy) or select the entity that appears first in the text (*first-loc* strategy). Therefore, four experiments were submitted combining both collection generated and both Geo-NER strategies. Table 1 shows the number of videos correctly predicted for different runs with the precision of 1, 5, 10, 50 and 100 kilometers.

During the processing of the 5091 test videos, our system found 89 videos (1,75%) without text in the *title* and *description* tags. Geo-NER did not find geographical entities in 3395 videos (66,68%). When Geo-NER did not find geographical entities in the *title* and *description* tags, for the two runs using *photos-meta* collection, 467 test videos (9,17%) were not georeferenced. On the other hand, for the two runs using *XML* collection, 798 test videos (15,67%) were not georeferenced. Analyzing the results obtained for the best precision (1 km), the experiment that combined *max-population* strategy and the *photos-meta* collection reached the best result with 187 videos correctly predicted (only 3,67% of total test videos). However, for the worst precision (100 km), the best result was reached by the experiment that combined *first-location* strategy and the *photos-meta* collection, obtaining 1085 videos correctly predicted (21,31% of total test videos).

The low number of videos with geographical entities detected by Geo-NER seems to be the main cause of poor system performance. This could be due to some problems identified during the processing of text labels from videos. Firstly, our Geo-NER module recognized words in capital letters that are not proper nouns or geographical entities because they exist as locations. Secondly, sometimes geographical entities are misspelled (eg *Indiapolis* instead of *Indianapolis*).

5. CONCLUSIONS AND FURTHER WORK

In this paper, we present a first approach for automatically placing videos uploaded in Flickr on the world map. Our system involves applying a geographical entity recognizer like Geo-NER to the textual annotations of each video. Geo-NER is based on different external knowledge resources, Wikipedia and GeoNames. If entities have not been detected, then we apply information retrieval using the metadata provided for development purposes. This metadata includes textual tags such as *title*, *description*, *location*, *keywords* or *geodata* about Flickr videos and images uploaded by users. The results obtained show that it is necessary to apply a more accurate recognition of geographical entities. An interesting issue would be to analyze more deeply how many videos, of which Geo-NER did not find geographical entities, really include geographical entities in its textual annotations. Another interesting task would be to study how many videos have geographical entities *implicitly*, that is, videos that refer to some location but do not contain any explicit geographical entity in its textual annotations. In these cases, apply any *tagger* like Geo-NER would be useless.

[5] `http://lucene.apache.org/`

6. REFERENCES

[1] Ahern, S., Naaman, M., Nair, R., and Yang, J. H.-I. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (2007), ACM, pp. 1–10.

[2] Amitay, E., Har'El, N., Sivan, R., and Soffer, A. Web-a-where: geotagging web content. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (2004), ACM, pp. 273–280.

[3] Crandall, D. J., Backstrom, L., Huttenlocher, D. P., and Kleinberg, J. M. Mapping the world's photos. In *WWW* (2009), J. Quemada, G. León, Y. S. Maarek, and W. Nejdl, Eds., ACM, pp. 761–770.

[4] Hays, J., and Efros, A. A. Im2gps: estimating geographic information from a single image. In *CVPR* (2008), IEEE Computer Society.

[5] Leidner, J. L. *Toponym Resolution in Text.* Universal-Publishers, 2008.

[6] Perea-Ortega, J. M., Montejo-Ráez, A., Martínez-Santiago, F., and Ureña-López, L. A. Geo-NER: un reconocedor de entidades geográficas para inglés basado en GeoNames y Wikipedia. *Sociedad Española para el Procesamiento del Lenguaje Natural 43* (2009), 129–136.

[7] Serdyukov, P., Murdock, V., and van Zwol, R. Placing flickr photos on a map. In *SIGIR* (2009), J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, and J. Zobel, Eds., ACM, pp. 484–491.