Toponym Resolution in Social Media

Neil Ireson and Fabio Ciravegna

University of Sheffield, UK

Abstract. Increasingly user-generated content is being utilised as a source of information, however each individual piece of content tends to contain low levels of information. In addition, such information tends to be informal and imperfect in nature; containing imprecise, subjective, ambiguous expressions. However the content does not have to be interpreted in isolation as it is linked, either explicitly or implicitly, to a network of interrelated content; it may be grouped or tagged with similar content, comments may be added by other users or it may be related to other content posted at the same time or by the same author or members of the author's social network. This paper generally examines how ambiguous concepts within user-generated content can be assigned a specific/formal meaning by considering the expanding context of the information, i.e. other information contained within directly or indirectly related content, and specifically considers the issue of toponym resolution of locations.

Keywords: Concept Disambiguation, Social networks, Information Extraction.

1 Introduction

The growth in the use of social media for sharing content (text, images or video) to other individuals who can be close personal associates or random strangers, is staggering. If the latest statistics from Facebook[1] are to be believed, around 7% of the world's population are active users and they spend on average 40 minutes per day on this one site. Whilst the value of this User-Generated Content (UGC) is being realised, utilising the information it contains poses a number of challenges. Contributions to social media sites (blogs, forums, Twitter, etc.) are conversational in nature and thus tend to be brief and informal, containing imprecise, subjective and ambiguous information. The provider of the content may make assumptions about the receivers' ability to interpret the meaning, despite the fact that the message (i.e. content and any associated metadata) may imperfectly represent their intended meaning. For example, incidental finding in a recent study on photo retrieval [1] indicated that people are unable to retrieve their own content due to their inconsistent descriptions.

One solution to this issue would be to facilitate the user in providing clear semantics defining any potential ambiguous concept they use. Recently a number

[1] `http://www.facebook.com/press/info.php?statistics`

P.F. Patel-Schneider et al. (Eds.): ISWC 2010, Part I, LNCS 6496, pp. 370–385, 2010.

of services (such as OpenCalais[2]) attempt to guide the user providing content to link concepts in their text to some URI, e.g. to Wikipedia or IMDB[3] articles. If such approaches are to be useful they must make suggestion which match the content provider's intentions. In order to determine the correct resolution of an ambiguous concept it is necessary to consider its context, whilst this context is most readily provided by the other information contained in the content (i.e. other text, image features, tags, etc.) the conversational nature of social media means that this information might well be limited and imperfect. However the interrelated nature of social media means that the disambiguation process may be able to use more distant but still related context, for example content posted around the same time or by the same user or by members of the user's social network.

In this paper we examine the use of this expanding context to resolve ambiguous concepts in social media and specifically consider the issue of toponym resolution, i.e. the allocation of specific geolocations to target location terms. Section 2 considers the related work; the disambiguation of both generic and location concepts in text and social media. Section 3 then outlines the methodology used to determine the expanding information context and discusses measurements for determining the degree of term ambiguity. Section 4 describes the experiment; the data used and generation of the disambiguation classification model. Section 5 presents the results and Section 6 discusses the short-comings of the work and how these might be addressed in the future. Section 7 then summarises the findings of the paper.

2 Related Work

The automatic disambiguation of concepts in social media has concentrated on the issue of ambiguous textual tags, this work can be broadly divided into two areas. The first approach disambiguates a target concept (tag) by creating clusters a frequently co-occurring tags, where each cluster is assumed to provide a separate meaning, defined by the tags it contains [2]. Such an approach has the advantage of being applicable to any tag, however as no specific meaning is assigned to each tag cluster it limits its usefulness and the ability to evaluate the approach. Although further processing has been used to assign a unique URI to the clusters based on the co-occurrence between cluster tags and terms found in an ontology [3], the work suffers from a limited evaluation of the techniques performance.

The second approach attempts to identify the correct meaning of tag given its use for a specific resource (i.e. image, web page). The co-occurring tags are used to provide context, these tags are compared with some tag-concept model to determine the most likely meaning. Angeletou [4] used WordNet [5] to identify ambiguous tags, however other work claims that WordNet tends to produce

[2] `http://www.opencalais.com/`
[3] Internet Movie Database: `http://www.imdb.com`

overly generic concepts [3]. Other approaches use purpose built tag-concept models based on Wikipedia/DBpedia [6,7,8].

The principal difficulty with these approaches is the issue of evaluating whether the disambiguation processes used actually assigns the correct meaning. All the studies perform post-experiment, human review of the results, and in general do not specify the nature of the evaluation (i.e. number of reviewers, inter-annotator agreement, etc.). The generation of "Gold-Standards" for the disambiguation of generic concepts in (multilingual) text has been undertaken by the SENSEVAL[4] evaluation exercises. However, this data is concerned with the identification of concepts in natural language, whilst in social media text, and particularly with tags, concepts have little or no grammatical context.

Toponym resolution has the advantage over general concept resolution that user-generated, gold standard data is available, and especially in social media data. This is due to the ability to, and interest in, geotagging UGC. In addition, in a limited context it is highly likely that a given location will have only one meaning [9], a hypothesis which is shown to be true for the data used in this experiment in respect to a given user context. A number of researchers have examined the disambiguation of locations in text, for example; in news articles [10], in Wikipedia articles [11] and general Web pages [12,13,14,15]. The disambiguation processes generally combine a number of techniques, including; statistical likelihood (selecting the most probable location), textual context (considering the surrounding text of a location) and co-referent locations within the document and, for web pages, hypertext-linked documents. The use of co-occurring locations to provide disambiguating context is stressed as a key technique which generally provides high precision, but low recall, when compared with other techniques. This is due to the requirement for related locations to occur, which may not be satisfied for a given document. One of the key issues in this technique is deriving a function to determine how co-occurrence of locations affects the disambiguation. Most frequently this involves using some heuristics to propagate toponym likelihood (based on location similarity, i.e. the spatial distance or the relative distance in some location taxonomy, between co-occurring locations and possible toponym resolutions), but more recent work forms a feature vector based on the co-occurrences and uses machine learning to calculate the most predictive function [11].

Crandall, et al. [16] combine image features and temporal context to geolocate Flickr images, and in their conclusion they indicate the potential advantage which could be derived from also considering social context. Davis, et al. [17] explore the combination of user and temporal context to determine the location of an uploaded photograph, but unfortunately do not provide enough description of the experimental results to determine the relative effects of these two contexts.

The recent work by Serdyukov, et al. [18], examines the issue of geolocating Flickr photos using the associated tags. Although their work does use the GeoNames[5] database to boost the importance of location names, the predictive model

[4] http://www.senseval.org/

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

incorporates all the photo tags. Their aim is not specifically toponym resolution, instead they attempt to calculate the actual latitude/longitude of a photo. The resultant model provides an association between a tag (or set of tags) and a location. This is similar to previous work looking at Flickr data to determine the location (and event) related semantics of a tag [19], i.e. the degree to which a given tag could be associated with a given location, and Wang et al.'s work on finding the relationships between news/blog tags and countries [20].

Weinberger, et al. [21] explore the general issue of tag ambiguity. The work defines ambiguity in terms of the probability of observing a given tag in a given context (i.e. set of tags). They then determine the two tags, if added to the set of tags, that give rise to maximally different probability distributions. For example the tags "UK" AND "MA" significantly effect the probability distributions of the tag "Cambridge". It is interesting to note that the research, which considers tag from 100 million images, indicates that 16% of tag ambiguity is explained by other geographic metadata. This emphasises the importance of determining the correct location associated with tags not only to geolocate UGC but also to disambiguate other tags. Indeed there has been work exploring the use of the known location (and time) of UGC to build a recommendation classifier to suggest other tags to the user [22,23].

Perhaps a surprising feature of the previous work on toponym resolution in social media, is that the social context has generally not been exploited. User models have been utilised in the disambiguation of general tags [6], and recent work, on tag recommendations [24] and determining the quality of reviews [25], have demonstrate that using social contextual information, i.e. information relating to a user and their social network, can help improve prediction especially overcoming the issue of data sparsity. The work in this paper examines how such information can be applied to improve the toponym resolution process.

3 Methodology

This section describes the techniques applied in toponym resolution (and applicable to concept disambiguation in general) and how social context can be used to improve performance.

3.1 Information Context

Similar to previous work the surrounding context is used to disambiguate the target concept. This context is provided by the tags associated with the UGC or the users themselves. There are both advantages and disadvantages in using tags over actual textual content; in text you can exploit grammatical structures and term proximity, whilst tags are, in effect, a "bag-of-words". However tags are intended to provide an overall description of the content so, if efficacious tagging is performed, tags should provide a valuable source of descriptive information. However not all tags will be equally informative and the degree of relevance of a given tag will depend upon the application. For toponym resolution the target concepts (tags) are locations and therefore it is necessary to

determine those tags which influence the location description of content. The approach adopted was to limit the tags by only considering location names, whilst there are a number of freely-available geographic resources, Yahoo! GeoPlanet[6] was selected due to it providing a semantically structured lexical database. Its specified aim is to provide "geo-referencing data on the Internet", which it does by providing a common naming convention (each location is allocated a URI in the form of a Where-On-Earth Identifier (WOEID)) and a framework or taxonomy describing the relative geography of these locations. The version[7] used contains over 5 million locations/toponyms, but more importantly the data provides an analogous structure to that found in general concept resources, such as WordNet. Each location is a node in a hierarchy from which it is possible to determine the parental locations (hypernyms) which contain the location, the child locations (hyponyms) that the location contains and also locations that share the same parent (coordinate terms). In addition it is possible to extract neighbouring locations, which are coordinate terms which are adjacent to the target location. A further attractive feature of Yahoo! GeoPlanet is the work on namespace concordance[8] which maps between the WOEID and a variety of other namespaces (e.g. location identifiers from; Geonames[9], OpenStreetMap[10], Wikipedia[11]). This means that it become possible to link content identified by a WOEID to information from multifarious providers including that from the Linked Data community.

The specific application scenario is a classic Information Retrieval problem whereby a user wishes to retrieve all the UGC which relates to a given instance of a concept, e.g. Sheffield, South Yorkshire, UK (WOEID:34503). The user can apply three strategies to retrieve the desired information:

1. Query for the ambiguous term and sift through the results. The effectiveness of this strategy is dependent upon the likelihood of content being tagged with the desired instance. If the user is looking for an obscure location, i.e. one which is relatively infrequently tagged, or the search term is highly ambiguous then many of the results will be irrelevant. In addition, if the location can be tagged with several synonyms (e.g. New York, NY, Big Apple) then relevant results may be missed.
2. Rely on the user to have tagged the content with the actual location URI. Note that with location it is also possible to use the geocoding coordinates of the content, if they are provided, although this does not necessarily uniquely identify the location, for example a point location (latitude/longitude) may refer to the immediate surroundings or it may simply be the central point of some wider area, e.g. city, county, country.

[6] `http://developer.yahoo.com/geo/geoplanet/`
[7] Version 7.5.1 released 2010-06-03
[8] `http://developer.yahoo.com/geo/geoplanet/guide/api-reference.html#api-concordance`
[9] `http://www.geonames.org`
[10] `http://www.openstreetmap.org`
[11] `http://en.wikipedia.org`

3. Form a query which is likely to return content relating to the desired instance. This strategy is reliant on the user's ability to form the complex query and the content being tagged with the disambiguating information. Weinberger, et al.'s [21] work on determining the most disambiguating tag would be relevant to directing the content provider tagging by suggesting the most effective tags.

The approach adopted, in effect, applies the third strategy by automatically constructing the complex query. In practice the process is applied offline allocating a location URI to every occurrence of a target location tag by considering all the co-occurring related location name tags: used to tag content, used by the tagging user and used by the users in their social network. Note that is is also possible to allocate a "non-location" URI, indicating that the target location term does not relate to any of the possible toponyms. Thus, information context is provided by a vector of related term frequencies:

$$IC = freq(T)_1, freq(T)_2, \ldots, freq(T)_l \tag{1}$$

and the meaning of any given term is provided by some function combining all the term's information contexts:

$$M(T) = f(IC_1, IC_2, \ldots, IC_n) \tag{2}$$

3.2 Ambiguity

The traditional types of ambiguity include lexical, syntactic, semantic, and pragmatic ambiguity (for a detailed discussion of ambiguity in natural language see [26]). This current work is only concerned with lexical ambiguity, that is where a term (i.e. a text string) has several different meanings. Lexical ambiguity can be subdivided into homonymy and polysemy. Homonymy occurs when a term can have a number of unrelated meanings, whilst polysemy occurs when a term has several related meanings. However this distinction is subtle and it is often unclear which type of ambiguity to apply, and is not considered in this work.

Mich [27] provides two measures for lexical ambiguity:

lexical ambiguity of a term T:

$$a(T) = \text{the number of meanings of T} \tag{3}$$

frequency-weighted lexical ambiguity of a term T:

$$a^*(T) = \sum_{i=1}^{a(T)} log_2 freq(M_i(T)), \tag{4}$$

where $M_i(T)$ is the i^{th} meaning of T, and freq(m) is the observed frequency of that meaning.

The meanings are provided by some lexical resource (e.g. WordNet) and the weighted function is calculated from the frequency of occurrences of meanings found in some text corpus. However the use of frequency seems erroneous; as a frequently used term with a single meaning is still deemed ambiguous. A preferred measure would be to use Shannon's information entropy, which more accurately measures the degree to which the occurrence of a term determines its meaning.

$$H(T) = -\sum_{i=1}^{a(T)} P(M_i(T)) log_2 P(M_i(T)), \tag{5}$$

where $P(M_i(T))$ is the probability of observing the i^{th} meaning of T.

Whilst a term may appear to be highly ambiguous due to a multitude of possible meanings (i.e. a high value according to Equation 3), in a given usage context only a limited set of those possibilities may be likely to occur. For example whilst there are 54 possible toponym resolutions of the location name Cambridge, any given user is only likely to refer to a single one of those possibilities. Although users may be unlikely to refer to multiple toponyms with the same name, they may use the term for meanings other than location names. For example, the term Barry can be associated with: 14 distinct toponyms, a common first name and can be used to describe a particular striped pattern in heraldry. Thus an individual user may use one of the non-geographic meanings in addition to using a single locational meaning for a given term. In the experiments report below the relative effect of term ambiguity on performance is considered.

4 Experiment

4.1 Data

The experiments were performed using Flickr data, a summary of the data is provided in Table 1. Three location areas were chosen; Cambridge, including Ely, Newmarket and Haverhill (as a classical example of an ambiguous location); Sheffield, including, Chesterfield, Barnsley, Hope Valley and Rotherham (for which an accurate local geographic database is available which can be used to assess the quality of the Geoplanet database); and Cardiff, including Barry, Ferndale, Sully, Penarth, Porth, Bridgend, Aberdare, Mountain Ash, Pentre, Cowbridge (which offers a number of highly non-ambiguous location names, and location names which are ambiguous due to also being common terms, namely Barry, Sully and Mountain Ash). These 20 target location names can be resolved into 268 toponyms. In total 1,143,529 photos were tagged with at least one of the these terms (after removing duplicates), of which 123,124 (10.8%) have an associated geolocation (latitude/longitude), these were uploaded by 12,326 users (approximately 10 photos per user). These geolocated photos are used to provide the gold standard.

The geolocated photos contain 165,389 target location name tags, note that each photo must contain at least one target location tag to be retrieved. The

users' 580,296 contacts produce 1,140,668 target location tags and the contacts' 5,700,749 contacts produce 3,998,763 target location tags. Whilst all the collected data was limited to an upload date before the end of 2009, all the contact and tag values are up-to-date at the time of retrieve (March 2010).

Each geolocation was then assigned to its nearest toponym, or, if it is greater than 30km away from any toponym it was assigned a null (i.e. non-location) value, this resulted in 99,215 photos assigned to toponyms and 23,909 "other" non-location meanings. When compared to Overell's [28] work on geolocation of Wikipedia articles, where the data contained 1,395 locations and 7,660 non-locations, the Flickr data contains over 22 times the proportion of location to non-location references. This may well be due to the fact tags are less likely to contain proper names (e.g. Person and Organisation names) when compared to free-text, as they are intended as a generic label. It is worth noting that in the experimental data for a given user all the occurrences of a specific target location term (e.g. Sheffield) resolve to a single toponym (e.g. WOEID:34503). However 1,229 (10%) of the users use the same term for both location and non-location meanings.

Table 1 shows the number of photos and users for each target location, note that the row values are not mutually exclusive, as a single document can contain to multiple location tags, therefore the totals are less than the sum of the rows. The final three columns provide measures for the term ambiguity, the first column, *Num*, gives the total number of meanings (toponyms) provided by the lexical database. The next two columns give the information entropy measures, computed from Equation 5, for the term ambiguity, the *Location* column considers the ambiguity with respect to the toponyms, whilst the *Term* column also includes the occurrence of non-location meanings. In general the inclusion of the non-location meaning increases the term ambiguity, however for the term Barry it is reduced due to the fact 85.1% of the occurrences of Barry refer to non-location meanings.

4.2 Classification

In order to resolve the location names it is necessary to determine their contextual information. As stated above this is provided by the co-occurrence of related location names, which are gleaned from the Yahoo! Geoplanet API. For each toponym the related location names are determined by: their *ancestors* (hypernyms), *children* (hyponyms) and *neighbours* (adjacent coordinate terms). However, whilst all the 268 toponyms have ancestors, only 36 possess children and 203 possess neighbours. In general the larger and more populous locations have a highly number of children, this is in part due to the fact such locations actual contain more child locations and also possible due to them being more accurately represented in the Geoplanet data. As a relatively accurate resource was available for the Sheffield area this was used to provide a basic analysis of the Geoplanet data. For Sheffield Geoplanet provides 43 child (suburbs) locations, whilst the more accurate resource provides 99 possibilities. In addition two of the suburb names provided by Geoplanet have incorrect spellings. Whilst

Table 1. Summary of location term data (number of photos, users and term ambiguity)

Location Name	Photos		Users		Ambiguity		
	All	Geo	All	Geo	Num	Location	Term
Cambridge	159969	29467	11881	2200	54	1.408	1.574
Ely	5953	1515	1608	301	13	1.388	1.852
Newmarket	4940	1020	664	154	16	2.135	2.384
Haverhill	3637	210	286	43	7	1.378	1.670
Cardiff	255012	36546	14337	2080	19	0.141	0.389
Barry	225629	29503	39337	3588	14	1.559	0.839
Ferndale	29722	6795	1953	299	30	1.515	1.801
Sully	26905	4450	6347	718	10	1.275	1.394
Penarth	12980	2652	1011	212	2	0.000	0.068
Porth	10060	2284	1785	384	2	0.392	1.154
Bridgend	5626	1109	654	140	14	0.857	1.109
Aberdare	5528	527	394	64	2	0.105	0.909
Mountain Ash	4222	392	1923	257	2	0.000	0.236
Pentre	1657	287	454	93	6	1.413	1.526
Cowbridge	1195	224	184	43	2	0.060	0.354
Sheffield	290253	39368	13424	2015	26	0.209	0.717
Chesterfield	40799	5907	4137	591	30	1.812	2.056
Barnsley	29589	6824	1460	240	7	0.022	0.554
Hope Valley	15692	1216	981	198	10	1.537	1.584
Rotherham	13970	2068	971	170	2	0.007	0.460
Total	1143529	123124	96109	12326	268		

such missing and erroneous data will adversely effect the absolute performance of the disambiguation process, the aim of the current work is to examine the relative performance of using an expanding context, rather than maximise the performance on the given data.

The co-occurrence between each of the 20 target location names and their related (i.e. ancestor, child and neighbour) locations is calculated. The document context is provided by all the related tags assigned to the photo. The user context is provided by all the related tags added by the user who uploads the photo, these tags are weighted by their frequency. The (uploading) user contacts' context is provided by all the related location tags added by the contacts, with tags weighted by the number of contacts who have used that tag. Similarly each of the contacts' contacts' tag usage provides further context. Although tags can be assigned any user, the vast majority of tags are provided by the uploading user, of the 8,193,877 location tags observed in the data only 23,534 (0.28%) are provided by other users. Thus four experiments are performed:

D : using only the related tags in the immediate document (photo) context
U : as D, including all the (uploading) user related tags as context

C : as U, including all the (uploading) user contacts' related tags as context
CC : as C, including all the (uploading) user contacts' contacts' related tags as context

For each experiment the set of co-occurring related location name frequencies provides a feature vector, from which a classification model is constructed. A Support Vector Machine (SVM) classifier used is (LibSVM [29]), applying a Radial-Basis Function kernel. For each experiment the feature vector values are normalised (between [0,1]) and a ten-fold cross-validation was performed. The photos uploaded by a specific user are placed in a single fold to prevent the classifier learning a user specific rather than generic classification model. For each fold the cost parameter was optimised using a three-fold cross-validation experiment on the training data. Note that along with the possible toponyms associated to a location term the classifier also learns to predict non-location references.

5 Results

Table 2 provides an overview of the experimental results, showing the location names, ordered according to increasing term ambiguity and the f-measure for the four experiments. Note that the reported f-measure for the locations is the micro-average of all the classes (toponyms), calculated by summing the one-versus-all matrices, as a result precision equals recall equals f-measure. The final row provides the macro-average of these micro averages.

The current approaches to concept disambiguation tend to rely on related information found solely within the context of the document, shown by the results in the second column. From these results it can be seen that including information from the creator of the content can significantly improve the disambiguation (paired t-test confidence <0.004), in addition including information from their social network contacts does produce some advantage but not highly significant (paired t-test confidence <0.42), whilst including information from the contacts' social network produces a slight detrimental effect over just using contacts' information. This trend can be observed in the macro-average values in the final row.

Note that the three location names where solely using the document context produced the best results have by far the three highest proportions of non-location meanings (i.e. Mountain Ash (0.961), Barry (0.851) and Sully (0.643)). Therefore the performance of such disambiguation techniques, which rely on the co-occurrence of related location names, are likely to be adversely affected by the presence of a significant proportion of non-location meanings.

Figure 1 and 2 examine the relationship between disambiguation performance and term ambiguity for the four experiments. Figure 1 shows the actual performance which indicates the expected decrease in performance with increasing ambiguity, this effect can be more clearly in Figure 2 which shows the trend lines for the data. The graphs indicate that for the experiments including user

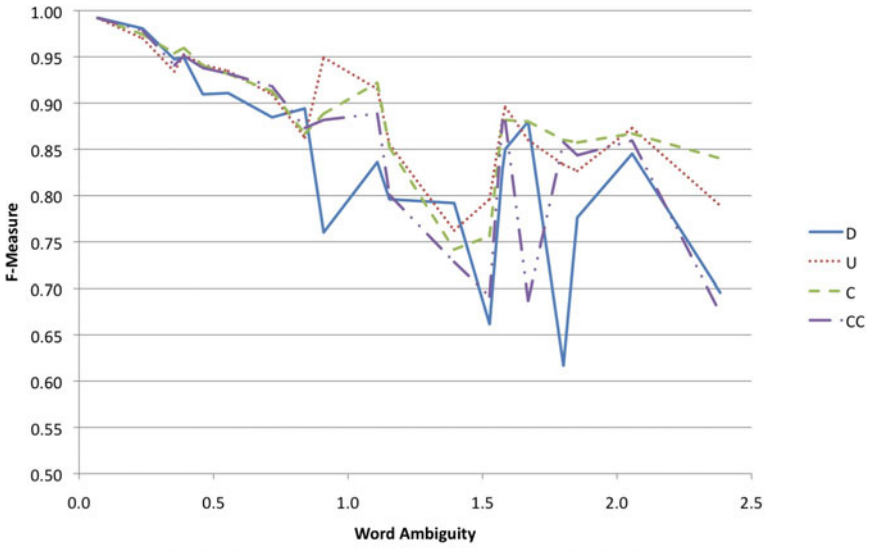

Fig. 1. Performance in relation to Word Ambiguity

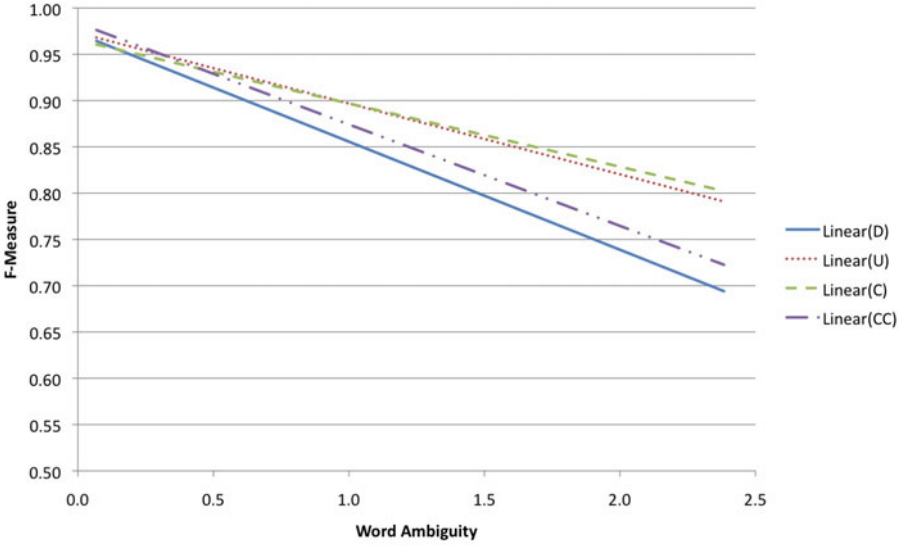

Fig. 2. Linear Trend in Performance in relation to Word Ambiguity

(U) and their social network contacts (C) as context the fall in performance is less sensitive to increase in term ambiguity.

Finally the overall results are depicted in Figure 3, which shows a generalised Precision-Recall curve for all classes (toponyms). The curve is formed using

Table 2. Performance measure for expanding social context

Location Name	Term Ambiguity	Performance (F-Measure)			
		Document	User	Contacts	Contacts' Contacts
Penarth	0.068	**0.992**	**0.992**	**0.992**	**0.992**
MountainAsh	0.236	**0.981**	0.970	0.974	0.979
Cowbridge	0.354	0.947	0.934	**0.954**	0.941
Cardiff	0.389	0.949	0.952	**0.960**	0.951
Rotherham	0.460	0.910	**0.941**	**0.941**	0.938
Barnsley	0.554	0.911	**0.935**	0.931	0.932
Sheffield	0.717	0.885	0.909	0.913	**0.918**
Barry	0.839	**0.894**	0.863	0.867	0.873
Aberdare	0.909	0.760	**0.949**	0.889	0.882
Bridgend	1.109	0.836	0.915	**0.922**	0.889
Porth	1.154	0.796	**0.856**	0.852	0.801
Sully	1.394	**0.792**	0.762	0.742	0.728
Pentre	1.526	0.662	**0.796**	0.756	0.692
Cambridge	1.574	0.828	**0.882**	0.879	0.880
HopeValley	1.584	0.850	**0.896**	0.882	0.880
Haverhill	1.670	0.880	0.860	**0.880**	0.687
Ferndale	1.801	0.617	0.833	**0.860**	0.858
Ely	1.852	0.776	0.827	**0.857**	0.844
Chesterfield	2.056	0.845	**0.873**	0.867	0.859
Newmarket	2.384	0.696	0.789	**0.840**	0.673
Macro-Average		0.832	0.891	**0.892**	0.860

a similar generalisation approach as outlined by Hand and Till [30] for ROC curves, whereby each of the individual Precision-Recall curves are combined with a weight according to the number of instances they represent, and missing points on each curve are calculate by linear interpolation.

The curve indicates that at low-recall values, where only highly probable instances are classified, considering more proximate contextual information (i.e. in the document rather than user tags, or user rather than contact tags) produces higher precision, which would be intuitively expected. However as recall increases utilising more distant context becomes more beneficial.

6 Discussion

The experiment and results described above indicate the importance of considering the user context when disambiguating location terms used to tag their content. In addition a potential benefit in considering the information provided by the user's social network contacts is shown. However the major limitation of the work is that it only uses a single data source, namely Flickr. Within this single domain an attempt was made to apply the techniques employed to a wide variety of concept types, varying the levels of term ambiguity and contextual

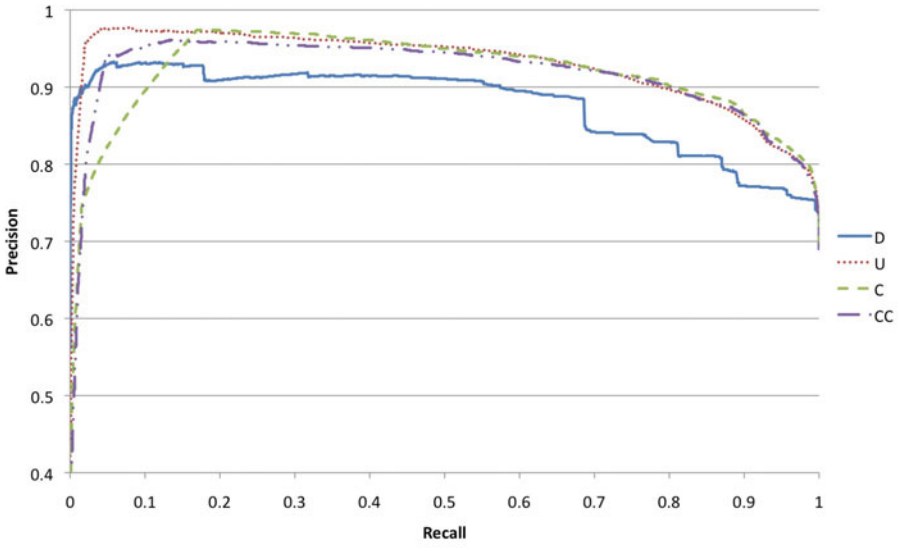

Fig. 3. Precision-Recall Curve

information available, and analyse the resultant performance. However drawing any conclusions outside the domain of toponym resolution in photo sharing location tags must be made with care, in particular for concepts where the user is more likely to use multiple meanings. Recently Twitter[12] have introduced the ability to geotag user content, which could be used to provide a comparison for toponym resolution in a different form of social media. For other, more generic concept types, the issue of creating an experimental gold standard to provide an objective evaluation needs to be addressed.

Although the experimental data described above provides a large number of labelled instances, a number of assumptions have been made. It is presumed that the user has accurately geocoded their content and any erroneous geocoding will provide a low level of noise, not significantly affecting the results. Although the experiment was mainly concerned with determining the relative probability of resolving a location term to competing toponyms, the limit of 30km set to determine that a geocoded photo was related to a given toponym is arbitrary and in practice this limit should be related to the toponym concept type, i.e. country, county, city, town, suburb, etc. In addition it should be noted that in their usage location terms do not have strict boundaries and the allocation of a given point to a toponym depends upon user context [31].

Whilst the main performance variability can be explained by the ambiguity of a given term (shown in the Figures 1 and 2), there is certain variation from the general trend. Other variables which potentially influence performance are the number of documents/users available for each target location, the number/type of related location tags (children, neighbours, ancestors) provided by

[12] `http://twitter.com/`

the GeoPlanet resource and the number of non-location references for each target location name. In addition, whilst Yahoo! Geoplanet was used as a semantically structured lexical database to provide a set of terms related to a concept, the resource is shown to contain missing and erroneous information. Further research should examine how such data and resource variables and imperfections influence the disambiguation process. The recent work on linking the various namespaces used to identify geolocations means it is possible to combine the information contained in these linked resources.

In this experiment related information is provided by the user and their social context, however a fairly naïve approach was adopted whereby all the user's information is assumed to relate equally to all their content. Similarly all the user's social network is assumed to have an equal impact. A more realistic approach would be to consider a temporal dimension, where information relatedness is dependent on temporal proximity, which previous work has shown to be effective [16]. If it is possible to determine the relative strength of social ties, e.g. with some measure of the degree of mutual interaction, this may also prove significant in determining the relative impact of information provided by the user's contacts. While the experiments showed that considering the information from the user's contacts' social network is detrimental to performance, considering the impact of social ties may allow information to be utilised from more distance social context.

7 Conclusion

This paper considers the issue of disambiguating concepts in social media. The approach adopted links the location concepts, found in user-generated content, to a URI defining its intended meaning; enabling the content to be retrieved according to a specific semantic query. Due to the nature of social media, the content containing the ambiguous concept may possess limited contextual information with which to disambiguate, however the interrelationship between content and users in social media means it is possible to exploit more distant, related contextual information. The paper shows the importance of considering user context when disambiguating the location terms they use to describe their content, and indicates that this is more important for terms with a higher degree of ambiguity. The information provided by a user's social network contacts can also provide some advantage although further work is required to determine if a more sensitive consideration of context, i.e. considering a temporal aspect or the strength of social ties, might improve the significance of using such social context.

Acknowledgments

This work has been supported by the European Commission as part of the WeKnowIt project (FP7-215453).

References

1. Whittaker, S., Bergman, O., Clough, P.: Easy on that trigger dad: a study of long term family photo retrieval. Personal Ubiquitous Comput 14(1), 31–43 (2010)
2. Yeung, C.m.A., Gibbins, N., Shadbolt, N.: Tag meaning disambiguation through analysis of tripartite structure of folksonomies. In: Web Intelligence/IAT Workshops, pp. 3–6. IEEE, Los Alamitos (2007)
3. Specia, L., Motta, E.: Integrating folksonomies with the semantic web. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 624–639. Springer, Heidelberg (2007)
4. Angeletou, S.: Semantic enrichment of folksonomy tagspaces. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 889–894. Springer, Heidelberg (2008)
5. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
6. Tesconi, M., Ronzano, F., Marchetti, A., Minutoli, S.: Semantify del.icio.us: Automatically turn your tags into senses. In: Social Data on the Web (2008)
7. Garcia, A., Szomszor, M., Alani, H., Corcho, O.: Preliminary results in tag disambiguation using dbpedia. In: Knowledge Capture (K-Cap 2009) - First International Workshop on Collective Knowledge Capturing and Representation - CKCaR 2009 (September 2009)
8. Overell, S., Sigurbjörnsson, B., van Zwol, R.: Classifying tags using open content resources. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM 2009, pp. 64–73. ACM, New York (2009)
9. Yarowsky, D.: One sense per collocation. In: Proceedings of the workshop on Human Language Technology, Morristown, NJ, USA, Association for Computational Linguistics, HLT 1993, pp. 266–271 (1993)
10. Garbin, E., Mani, I.: Disambiguating toponyms in news. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT 2005, Morristown, NJ, USA, Association for Computational Linguistics, pp. 363–370 (2005)
11. Overell, S., Rüger, S.: Using co-occurrence models for placename disambiguation. International Journal of Geographical Information Science 22, 265–287 (2008)
12. Ding, J., Gravano, L., Shivakumar, N.: Computing geographical scopes of web resources. In: Proceedings of the 26th International Conference on Very Large Data Bases, VLDB 2000, pp. 545–556. Morgan Kaufmann Publishers Inc., San Francisco (2000)
13. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: geotagging web content. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2004, pp. 273–280. ACM, New York (2004)
14. Clough, P., Sanderson, M., Joho, H.: Extraction of semantic annotations from textual web pages. Deliverable D15 6201, EU Project: SPIRIT (2004)
15. Zong, W., Wu, D., Sun, A., Lim, E.P., Goh, D.H.L.: On assigning place names to geography related web pages. In: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital libraries, JCDL 2005, pp. 354–362. ACM, New York (2005)
16. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: Proceedings of the 18th International Conference on World Wide Web, WWW 2009, pp. 761–770. ACM, New York (2009)

17. Davis, M., King, S., Good, N., Sarvas, R.: From context to content: leveraging context to infer media metadata. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, MULTIMEDIA 2004, pp. 188–195. ACM, New York (2004)
18. Serdyukov, P., Murdock, V., van Zwol, R.: Placing flickr photos on a map. In: Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, pp. 484–491. ACM, New York (2009)
19. Rattenbury, T., Naaman, M.: Methods for extracting place semantics from flickr tags. ACM Trans. Web 3(1), 1–30 (2009)
20. Wang, C., Wang, J., Xie, X., Ma, W.Y.: Mining geographic knowledge using location aware topic model. In: Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, GIR 2007, pp. 65–70. ACM, New York (2007)
21. Weinberger, K.Q., Slaney, M., Van Zwol, R.: Resolving tag ambiguity. In: Proceeding of the 16th ACM International Conference on Multimedia, MM 2008, pp. 111–120. ACM, New York (2008)
22. Naaman, M., Paepcke, A., Garcia-Molina, H.: From where to what: Metadata sharing for digital photographs with geographic coordinates. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds.) CoopIS 2003, DOA 2003, and ODBASE 2003. LNCS, vol. 2888, pp. 196–217. Springer, Heidelberg (2003)
23. Sarin, S., Nagahashi, T., Miyosawa, T., Kameyama, W.: On the design and exploitation of user's personal and public information for semantic personal digital photograph annotation. Adv. MultiMedia 2008(2), 1–16 (2008)
24. Rae, A., Sigurbjrnsson, B., van Zwol, R.: Improving tag recommendation using social networks. In: RIAO 2010, Paris, France (2010)
25. Lu, Y., Tsaparas, P., Ntoulas, A., Polanyi, L.: Exploiting social context for review quality prediction. In: 19th International World Wide Web Conference, WWW 2010 (April 2010)
26. Ceccato, M., Kiyavitskaya, N., Zeni, N., Mich, L., Berry, D.M.: Ambiguity identification and measurement in natural language texts. Technical Report Technical Report DIT-04-111, Univeristy of Trento (December 2004)
27. Mich, L.: On the use of ambiguity measures in requirements analysis. In: Proceedings of the 6th International Workshop on Applications of Natural Language to Information Systems, NLDB 2001, pp. 143–152. GI (2001)
28. Overell, S.: Geographic Information Retrieval: Classification, Disambiguation and Modelling. PhD thesis, Imperial College London (2009)
29. chung Chang, C., Lin, C.J.: Libsvm: a library for support vector machines (2001) Software available at, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
30. Hand, D.J., Till, R.J.: A simple generalisation of the area under the roc curve for multiple class classification problems. Mach. Learn. 45(2), 171–186 (2001)
31. Jones, C.B., Purves, R.S., Clough, P.D., Joho, H.: Modelling vague places with knowledge from the web. Int. J. Geogr. Inf. Sci. 22(10), 1045–1065 (2008)