A UTILITY CENTERED APPROACH FOR EVALUATING AND OPTIMIZING GEO-TAGGING

Albert Weichselbraun
Department of Computational Methods, Vienna University of Economics and Business, Vienna, Austria
albert.weichselbraun@wu.ac.at

Keywords: geo-tagging, quality assessment, evaluation, utility model, GeoNames

Abstract: Geo-tagging is the process of annotating a document with its geographic focus by extracting a *unique* locality that describes the geographic context of the document as a whole (Amitay et al., 2004). Accurate geographic annotations are crucial for geospatial applications such as Google Maps or the IDIOM Media Watch on Climate Change (Hubmann-Haidvogel et al., 2009), but many obstacles complicate the evaluation of such tags.

This paper introduces an approach for optimizing geo-tagging by applying the concept of utility from economic theory to tagging results. Computing utility scores for geo-tags allows a fine grained evaluation of the tagger's performance in regard to multiple dimensions specified in use case specific domain ontologies and provides means for addressing problems such as different scope and coverage of evaluation corpora.

The integration of external data sources and evaluation ontologies with user profiles ensures that the framework considers use case specific requirements. The presented model is instrumental in comparing different geo-tagging settings, evaluating the effect of design decisions, and customizing geo-tagging to a particular use cases.

1 INTRODUCTION

The vision of the Geospatial Web combines geographic data, Internet technology and social change. Geospatial applications such as the IDIOM Media Watch on Climate Change (Hubmann-Haidvogel et al., 2009) use geo-annotation services to refine Web pages and media articles with geographic tags. Geo-tagging is the process of assigning a *unique* geographic location to a document or text. In contrast to geographic named entity recognition or toponym resolution (Leidner, 2006) only *one* geographic location which describes the document's geography is extracted, even if multiple geographic references occur in the document.

Most approaches toward geo-tagging facilitate machine learning technologies, gazetteers, or a combination of both to identify geo-entities. The gazetteer's size and tuning parameters determine the geo-tagger's performance and its bias towards smaller geographic-entities or higher-level units. Choosing these parameters often involve trade-offs; improvements in one particular area do not necessarily yield better results in other areas.

For instance, increasing the gazetteer's size increases the number of detected geographic entities but comes at the cost of a higher probability of ambiguities. Gazetteer entries such as *Fritz/at*, *Mobile/Alabama/us*, *Reading/uk* challenge the tagger's capability to distinguish geographic entities from common terms without a geographic meaning. Therefore, a framework which monitors the effect of design decisions on the tagger's performance and yields comparable performance metrics is essential for designing and evaluating geo-taggers.

Clough and Sandner (Clough and Sanderson, 2004) point out the importance of comparative evaluations of geo-tagging as stimuli for academic

and industrial research. Leidner (Leidner, 2006) provides such an evaluation data set and describes the process of designing evaluation corpora. Nevertheless evaluating geographical data mining is still a tricky task. Martins et al. (Martins et al., 2005) elaborate on the challenges required to develop accurate methods for evaluating geographic tags, which include the creation of geographic ontologies, interfaces for geographic information retrieval, and the development of methods for ranking documents according to geographic relevance.

Providing a generic evaluation framework to compare geographic annotations is still a rather complex task. Parameters such as the gazetteer's scope, coverage, correctness, granularity, balance and richness of annotation influence the outcome of any evaluation experiment (Leidner, 2006). Therefore, even standardized evaluation corpora such as the one designed by Leidner require geo-taggers to use a fixed gazetteer to provide comparable results.

Studies show (Hersh et al., 2000; Allan et al., 2005) that information retrieval performance measures as for instance *recall* do not always correspond to adequate gains in actual user satisfaction (Turpin and Scholer, 2006). Work by Turpin and Hersh (Turpin and Hersh, 2001) suggests that improvements of information retrieval metrics do not necessarily translate into better user performance for specific search tasks.

Martins et al. (Martins et al., 2005) recommend to close the gap between performance metrics and user experience by performing user studies. Despite the additional effort required to implement such studies, work by Nielsen and Landauer (Nielsen and Landauer, 1993) suggests that approximately 80% of the described usability problems can be detected with only five users (Martins et al., 2005).

This work addresses the need for comparative evaluations and user participation by applying the concept of utility to geo-tagger evaluation metrics. Intra-personal settings translate tagging results into utility values and allow to measure the performance according to the user's specific needs.

The remainder of this paper is organized as follows. Section 2 elaborates on challenges faced in geo-tagging. Section 3 presents a blueprint for applying the concept of utility to geo-tagging and describes the process of deploying a geo-evaluation ontology. Section 4 demonstrates the usefulness of utility centered evaluations by comparing the utility based technique to conventional approaches. The paper closes with an outlook and draws conclusions in Section 5.

2 EVALUATING GEO-TAGS

Web pages often contain multiple references to geographic locations. State of the art geo-taggers facilitate these references to identify the site's geographic context and resolve ambiguities using the obtained context. A focus algorithm decides based on the identified geographic entities on the site's geography (Amitay et al., 2004). Tuning parameters determine the focus algorithm's behavior, such as whether it is biased toward higher-level geographic units (such as countries and continents) or prefers low-level entities such as cities or towns.

Biases make judging the tagger's performance difficult. An article about Wolfgang Amadeus Mozart, for example, contains one reference to Salzburg and two to Vienna - both cities in Austria. Depending on the focus algorithm's configuration, the page's geography might be set to (i) Salzburg (bias toward low-level geographic units), (ii) Austria (bias toward high-level geographic units), or (iii) Vienna (bias toward low-level geographic units with a large population).

The task of judging the value of a particular answer is far from trivial, because each possible solution has a certain *degree of correctness*. Work comparing results to a gold standard often fails to value these nuances.

This paper therefore suggests to apply the concept of utility, as found in economic theory, to the evaluation of geo-taggers. The geographies returned by the tagger are assessed based on preferences specified by the user along different ontological dimensions and get scored accordingly.

Maximizing utility instead of the number of correctly tagged documents, provides advantages in regard to: (i) *granularity* - the architecture even accounts for slight variations in the grade of "correctness" of the proposed geo-tags; (ii) *adaptability* - users can specify their individual utility profiles, providing the architect with means to assess the tagger's performance in accordance with the particular preferences of a user; and (iii) *holistic observability* - the geo-tagger's designer is no longer restricted to observe gains, but can consider costs in terms of computing power, storage, network traffic, and response times.

3 METHOD

Figure 1 outlines how the utility based approach uses ontologies to evaluate the geo-tagging performance. The framework compares the geo-tagger's

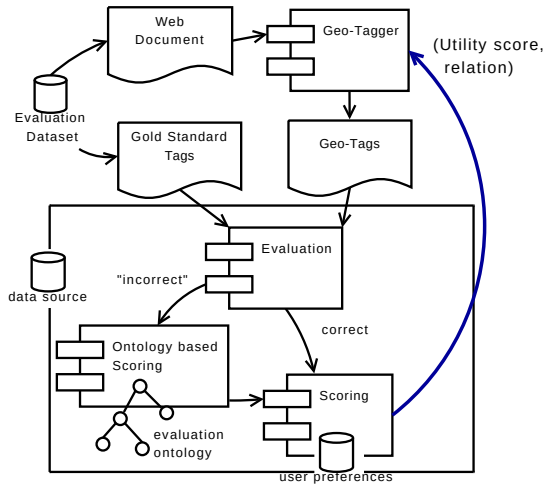

Figure 1: Ontology-based evaluation of geo-tags.

annotations with tags retrieved from a gold standard. Correct results yield the full score, incorrect results are evaluating using ontology based scoring which verifies whether the result is related to the correct answer in regard to the dimensions specified in the evaluation ontology and the extend of such a possible relation. Queries against the data source identify such ontological relationships between the computed and the correct tag, which are evaluated considering the answer's deviation from the correct answer and the user's preference settings.

3.1 Ontology and Data Source

The evaluation ontology specifies the ontological dimensions considered for the evaluation task. Object properties such as *x partOf y*, or *x isNeighbor y* specify the relations between the "correct" answer and its deviations.

The data source provides instance data covering the location entities identified by the tagger. It therefore allows querying pairs of objects to retrieve their relations in regard to the evaluation ontology. Data source and evaluation ontology are closely related. Depending on the use case and available resources a bottom-up (design the ontology according to an existing data source) or a top-down approach (design the ontology and create a fitting data source) will be chosen for the evaluation ontology's design. The ontology's object properties specify valid ontological dimensions for the evaluation process.

Existing ontologies containing geographical categories as for instance the one applied by David Warren and Fernando Pereira (Warren and Pereira, 1982) in the Chat-80 question-answering system may act as a template for such an evaluation ontology. This work uses a bottom-up approach based on the publicly available GeoNames database (geonames.org). GeoName's *place hierarchy web service* provides functions to determine an entry's children, siblings, hierarchy, and neighbors. Functions such as *findNearby* return streets, place names, postal codes, etc. for nearby locations, and auxiliary methods deliver annotations such as postal codes, Wikipedia entries, weather stations and observations for a given location. For a full list of the supported functions please refer to the GeoNames Web service documentation[1].

Due to the applied bottom-up approach the created ontology only considers relations derived from GeoName entries. Despite the ontology's general scope its application to other use cases might require refinements of the ontological constructs. The ontology supports standard properties such as *partOf*, *isNeighbor* and *sibling* relationships as well as data type properties assigning entities coordinates (*centerCoordinates*), an area (*totalArea*), and a population (*totalPopulation*), if applicable. The *contains* property helps distinguishing between geo-entities completely containing another entity (e.g. Europe *contains* Austria), and entities which are only partly contained by another entity (e.g., Russia is *partOf* Europe, but Europe does not *contain* it).

Combining the geo-evaluation ontology's knowledge with queries for ontological instances in the GeoNames database yields an effective framework for the evaluation of geographic tags. Queries alongside the ontological dimensions allow a fine grained assessment of the tagger's result including the extend to which "incorrect" tags contribute helpful information.

3.2 User Preferences

User preferences determine the translation of test results into utility scores. Equation 1 shows a utility function assuming linearly independent utility values.

$$u = \sum_{a_i \in S_A} f_{eval}(a_i) \quad (1)$$

[1] www.geonames.org/export/ws-overview.html

The utility equals to the sum of the utility gained by a answer set $S_A = \{a_1, a_2, \ldots, a_n\}$, which is evaluated using an evaluation function f_{eval}. To simplify the computation of the utility many current evaluation metrics only consider correct answers as useful ($f_{eval} = 1$ for correct answers and 0 otherwise).

Such approaches are too coarse to detect minor deteriorations in the tagger's performance, because the utility generated by a particular answer is highly use case and user specific. Thus designing geo-taggers requires more fine grained methods which consider the user's preferences and fine nuances of correctness.

The evaluation ontology outlines these nuances in terms of ontological dimensions and the user preferences address the issue of assigning use case specific weights to those dimensions. This approach adds the following Equation to evaluate partly correct answers:

$$f_{eval}(a_i) \quad = \quad \prod_{j=1}^{n} w_{d_j} \qquad (2)$$

with a user specific weight w_{d_j} $[0, 1]$ for deviations alongside the ontological dimension d_j. Identification of paths between the tag a_k and the correct answer a_k^* along the ontological dimensions yields one w_{d_j} for every movement. If no path between a_k and a_k^* exists, f_{eval} is set to zero, if multiple paths lead to a_k^* the framework applies resolving strategies such as (i) use the shortest path, (ii) maximize $\prod_{j=1}^{n} w_{d_j}$, or (iii) summarize the utility of all paths and use $f_{eval} = min(f_{eval}^{sum}, 1)$.

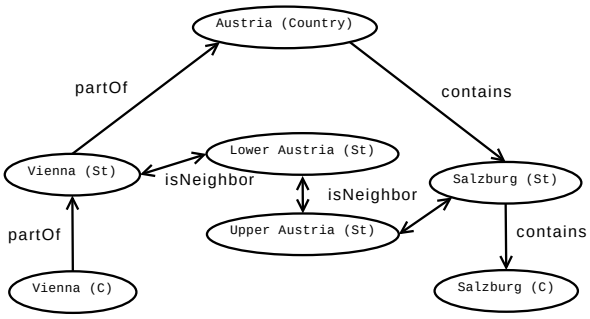

Figure 2: Evaluation of partially correct results.

Figure 2 demonstrates the application of the scoring procedure which facilitates an evaluation ontology designed for proximity based scoring. In the example the tagger provides the tag *Salzburg* instead of *Vienna*. Two paths lead to the correct answer: (i) Vienna (City) via *partOf* to Vienna (State) via *isNeighbor* to Lower Austria, Upper Austria and Salzburg (State) via *contains* to Salzburg (City), and (ii) Vienna (City) via *partOf* to Vienna (State) and Austria (Country) via *contains* to Salzburg (State) and Salzburg (City). Depending on the chosen resolution strategy f_{eval} equals to

$$(w_{partOf}(w_{isNeighbor})^3 w_{contains}) \quad \text{or}$$
$$((w_{partOf})^2 (w_{contains})^2).$$

3.3 Scoring

Many heuristics for the evaluation of geo-tags have emerged. Martins et al. (Martins et al., 2005) provide a number of possible measures for geographical relevance including (i) Euclidean distance, (ii) extend of overlap, (iii) topological distance as for instance adjacency, connectivity or hierarchical containment, and (iv) the similarity in semantic structures. This work proposes a hybrid approach considering Euclidean distance, hierarchical containment, and semantic structures as formalized in the evaluation ontology by computing similarity based on the number of correctly identified hierarchy levels and the distance between the correctly and incorrectly tagged entity.

At first a tagging result is followed along its hierarchical structure (compare Figure 3) until its geo-entity differs from the correct answer. The tagging utility u_c consists of a utility for the correctly identified hierarchical levels u_c^h and a utility assigned to deviations along the dimensions specified in the evaluation ontology u_c^o for partially correct entries:

$$u_c \quad = \quad u_c^h + u_c^o \qquad (3)$$
$$u_c^o \quad = \quad (1 - u_c^h) \cdot f_{eval} \qquad (4)$$

The algorithm computes u_c^o and u_c^h based on the number of geographic levels on which the results disagree:

$$u_c^h \quad = \quad \frac{|S_{correct} \cap S_{suggested}|}{max(|S_{correct}|, |S_{suggested}|)} \qquad (5)$$

Equal tags yield an u_c^h of one and therefore u_c^o of zero. Deviations between the tags lead to $u_c^h < 1$ and $u_c^o > 0$.

Equations 6 and 7 show how f_{eval} is composed when applying the distance centered evaluation. The idea of this method is to combine the information retrieved in terms of deviations alongside the ontological dimensions in the evaluation ontology with the *additional accuracy* retrieved from the "wrong" data based on the distance between the given and the correct location (d) in

$$\underbrace{at}_{u_c^h} \quad \underbrace{/Carinthia/Spittal/Heiligenblut}_{u_c^o} \quad \text{detected}$$
$$at \quad /National\ Park\ Hohe\ Tauern \quad \text{correct}$$

Figure 3: Scoring for hierarchical u_c^h entries and deviation alongside dimensions specified in the evaluation ontology u_c^o.

comparison to the expected distance (d_e) between two randomly selected points in a circular area as big as the area of the last correct item (A_{S_l}) in the tagging hierarchy:

$$d_e = E(d_{random}) = \frac{1}{3}\sqrt{A_{S_l}/\pi} \quad (6)$$

$$f_{eval}^d = max(0, (1 - \frac{d}{d_e}\prod_{i=1}^{n} w_{di})) \quad (7)$$

Summarizing the utility gained from the identified geographic entities yields the tagger's total utility for a particular tagging use case.

4 EVALUATION

To demonstrate the influence of user specific settings such as the gazetteer size or the tagger's scope on the geo-tagger's results, an evaluation facilitating 15 000 randomly selected articles from the Reuters corpus (trec.nist.gov/data/reuters/reuters.html) has been performed. The evaluation compares results obtained from the OpenCalais Web service (www.opencalais.com) and the geoLyzard-tagger used in the IDIOM Media Watch on Climate Change (www.ecoresearch.net/climate) with location reference data from the Reuters corpus. The Reuters corpus specifies the location on a fixed scope (country or political organization), while both other taggers determine the scope dynamically based on the document's content.

The experiment evaluates geo-tags according to four different criteria:

1. *verbatim correctness ($A \equiv B$)*: Both geo-tagger identify exactly the same geographic entity.
2. *more detailed specification ($A \sqsupseteq B$)*: The found location is an equal or a more detailed specification of the gold standard's entity (e.g. *eu/at/Salzburg* is more detailed than *eu/at*).
3. *more general specification ($A \sqsubseteq B$)*: The tagger returns an equal or more general specification of the gold standard's entity (e.g. *eu/it* is a more general specification than `eu/it/Florence`).
4. *more detailed or more general specification ($A \sqsupseteq B \vee A \sqsubseteq B$)*: The location satisfies either condition 2 or 3.

In contrast to the evaluation of a tag's verbatim correctness the other three test settings require domain knowledge as outlined in Section 3.

Table 1 summarizes the evaluation's results. Both tagger tend to deliver most of the data at a more fine grained scope than country level which leads to only around 20% of *verbatim* conformance with the gold-standard. Considering hierarchical data in the evaluation boosts the evaluation metric to approximately 75%. The rest of the deviations might be caused by (i) different configurations of the foci algorithms used in the taggers, (ii) by changes in the geopolitical situations as for instance the break-up of Yugoslavia into multiple countries, (iii) by missing geographic references in the original articles, and (iv) real misclassifications. Deviations due to different foci algorithms as well as changes in the geopolitical situation might be addressed by extended evaluation ontologies, supporting more complex relations between the geo-entities. In contrast, an evaluation of the test corpus and a manual inspection of the returned geo-tags is required to quantify the share of the latter two causes.

The experiment illustrates how the inclusion of domain knowledge improves the comparability of geo-tagging evaluation metrics. The presented evaluation only uses a subset of two relations (*partOf* and *contains*) from the ontology introduced in Seciton 3. Applying all relations available at GeoNames will yield even more accurate performance metrics. More sophisticated approaches might even implement geographic reasoning (e.g., through Voronoi polygons (Alani et al., 2001) or spatial indexes based on uniform grids (Riekert, 2002)).

5 OUTLOOK AND CONCLUSIONS

This work presented a utility-testing centered approach for optimizing geo-taggers. The contributions of this paper are (i) introducing a fine grained notion of *correctness* in terms of a tagging utility applicable to geo-tagging results, (ii) presenting an approach for the evaluation of geo-taggers, (iii) demonstrating the concrete imple-

Comparison	$=$	$A \sqsupseteq B$	$A \sqsubseteq B$	$A \sqsubseteq B \vee A \sqsupseteq A$
OpenCalais vs. Reuters	20.15 %	71.68 %	31.45 %	78.43 %
geoLyzard vs. Reuters	16.82 %	62.25 %	25.01 %	74.50 %
OpenCalais vs. geoLyzard	47.25 %	50.63 %	48.15 %	62.23 %

Table 1: Evaluation of geo-tags created by OpenCalais and geoLyzard.

mentation of such a framework by designing a geo-evaluation ontology customized to be used together with the GeoNames Web service, and (iv) evaluating the effect of ontological knowledge and external data on the evaluation metrics.

To Compare utility instead of geo-tags aids in overcoming obstacles such as different scopes, foci algorithms, granularity, and coverage. Conventional approaches which limit the tagger's scope or standardize the used gazetteer are not feasible to evaluate more sophisticated applications which provide tags at many different scopes according to the user's preferences. The notion of utility provides a very fine grained, user specific measure for the tagger's performance. Community efforts such as FreeBase and WikiDB provide a solid base for extending this method to other dimensions as outlined in Section 2. Considering query cost in evaluating the tagger's performance is another interesting research avenue (Weichselbraun, 2008).

Future research will transfer these techniques and results to more complex use cases and integrate multiple data sources.

REFERENCES

Alani, H., Jones, C. B., and Tudhope, D. (2001). Voronoi-based region approximation for geographical information retrieval with gazetteers. *International Journal of Geographical Information Science*, 15(4):287–306.

Allan, J., Carterette, B., and Lewis, J. (2005). When will information retrieval be "good enough"? In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 433–440, New York, NY, USA. ACM.

Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). 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*, pages 273–280, New York, NY, USA. ACM.

Clough, P. and Sanderson, M. (2004). A proposal for comparative evaluation of automatic annotation for geo-referenced documents. In *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*.

Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L., and Olson, D. (2000). Do batch and user evaluations give the same results? In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 17–24, New York, NY, USA. ACM.

Hubmann-Haidvogel, A., Scharl, A., and Weichselbraun, A. (2009). Multiple coordinated views for searching and navigating web content repositories. *Information Sciences*, 179(12):1813–1821.

Leidner, J. L. (2006). An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30:400–417.

Martins, B., Silva, M. J., and Chaves, M. S. (2005). Challenges and resources for evaluating geographical ir. In *GIR '05: Proceedings of the 2005 workshop on Geographic information retrieval*, pages 65–69, New York, NY, USA. ACM.

Nielsen, J. and Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems*, pages 206–213, Amsterdam, The Netherlands. ACM.

Riekert, W.-F. (2002). Automated retrieval of information in the internet by using thesauri and gazetteers as knowledge sources. *Journal of Universal Computer Science*, 8(6):581–590.

Turpin, A. H. and Hersh, W. (2001). Why batch and user evaluations do not give the same results. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 225–231, New York, NY, USA. ACM.

Turpin, A. H. and Scholer, F. (2006). User performance versus precision measures for simple search tasks. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11–18, New York, NY, USA. ACM.

Warren, D. H. D. and Pereira, F. C. N. (1982). An efficient easily adaptable system for interpreting natural language queries. *Computational Linguistics*, 8(3-4):110–122.

Weichselbraun, A. (2008). Strategies for optimizing querying third party resources in semantic web applications. In Cordeiro, J., Shishkov, B., Ranchordas, A., and Helfer, M., editors, *3rd International Conference on Software and Data Technologies*, pages 111–118, Porto, Portugal.