Things and Strings: Improving Place Name Disambiguation from Short Texts by Combining Entity Co-Occurrence with Topic Modeling

Yiting Ju[1], Benjamin Adams[2], Krzysztof Janowicz[1], Yingjie Hu[3], Bo Yan[1], and Grant McKenzie[4]

[1] STKO Lab, University of California, Santa Barbara, USA
[2] Centre for eResearch, The University of Auckland, New Zealand
[3] Department of Geography, University of Tennessee, Knoxville, USA
[4] Department of Geographical Sciences, University of Maryland, USA
yju@umail.ucsb.edu,b.adams@auckland.ac.nz,janowicz@ucsb.edu,
yhu21@utk.edu,boyan@umail.ucsb.edu,gmck@umd.edu

Abstract. Place name disambiguation is the task of correctly identifying a place from a set of places sharing a common name. It contributes to tasks such as knowledge extraction, query answering, geographic information retrieval, and automatic tagging. Disambiguation quality relies on the ability to correctly identify and interpret contextual clues, complicating the task for short texts. Here we propose a novel approach to the disambiguation of place names from short texts that integrates two models: entity co-occurrence and topic modeling. The first model uses Linked Data to identify related entities to improve disambiguation quality. The second model uses topic modeling to differentiate places based on the terms used to describe them. We evaluate our approach using a corpus of short texts, determine the suitable weight between models, and demonstrate that a combined model outperforms benchmark systems such as DBpedia Spotlight and Open Calais in terms of F1-score and Mean Reciprocal Rank.

Keywords: Place name disambiguation; natural language processing; LDA; Wikipedia; DBpedia; Linked Data

1 Introduction

Geographic knowledge extraction and management, geographic information retrieval, question answering, and exploratory search hold great promise for various application areas [19, 12, 2]. From intelligence and media analysis to socio-environmental studies and disaster response, there is demonstrated need to be able to build computational systems that can synthesize and understand human expressions of information about places and events occurring around the world [8]. Being able to correctly identify geographic references in the abundance of

unstructured textual information now available on the Web, in social media, and in other communication media is the first step to building tools for geographic analysis and discovery on these data. Place name, i.e., toponym, disambiguation is key to the comprehension of many texts as place names provide an important context required for the successful interpretation of text [13].

Similar to other named entities, including persons, organizations, and events, place names can be ambiguous. A single place name can be shared among multiple places. To give a concrete example, *Washington* is a place name for more than 43 populated places in the United States alone.[5] Although most of these Washingtons can be accurately located by adding the proper state name or county name, they are all simply referred to as *Washington* in daily conversations, (social) media, photo annotations, and so forth. Figure 1 depicts the distribution of the most common place names for U.S. cities, towns, villages, boroughs, and census-designated places. As shown on the map, these places are distributed across the U.S., indicating that the ambiguity of place names is a widespread phenomenon. It is worth noting that places which share a common name can be of the same or a different type, e.g., the *state* of Washington and the *city* of Washington, Pennsylvania. The situation is even more difficult on a global scale where place names may appear more than 100 times. For example, it takes merely a 45min car ride to get from Berlin to East London, both located in South Africa. Thus, it is important to devise effective computational approaches to address the disambiguation problem.

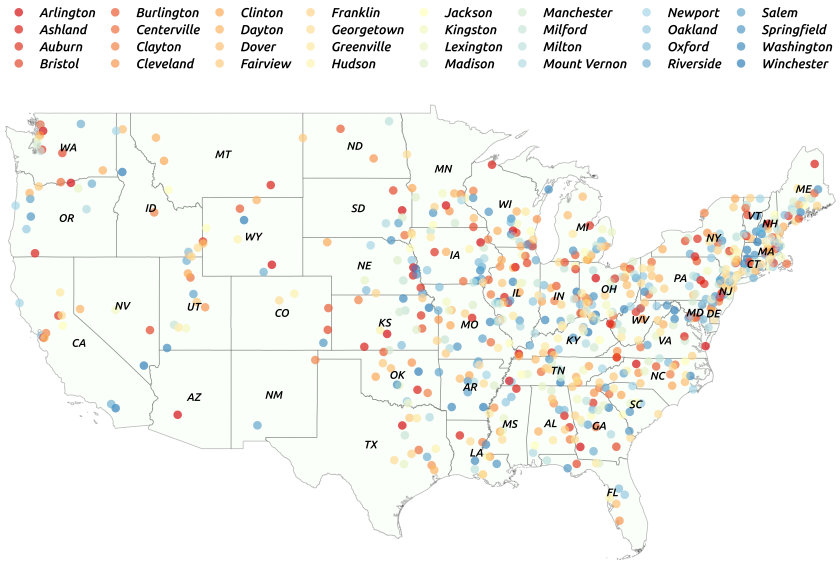

Fig. 1: Distribution of common place names in the US according to Wikipedia.

[5] https://en.wikipedia.org/wiki/Washington

Given the wide availability of digital gazetteers, i.e., place name dictionaries, such as GeoNames, the Getty Thesaurus of Geographic Names, the Alexandria Digital Library Gazetteer, and Google Places, we assume that the places to be disambiguated are known, i.e. that there is a candidate list of places for any given place name list. After all, unknown places cannot be disambiguated. Thus, we define the task of place name disambiguation as follows: *given a short text which contains a place name and given a list of candidate places that share this name, determine to which specific place the text refers.*

Humans are very good at detecting and interpreting contextual clues in texts to disambiguate place names. Thus, as extension of named entity recognition, place name disambiguation has been tackled using computational approaches that aim at utilizing these contextual clues as well [5, 7]. This context typically stems from the terms surrounding the place name under consideration. Typically, short texts from social media, news headlines (and abstracts), captions, and so forth, offer less contextual clues and thus negatively impact disambiguation quality. Consequently, new approaches have to be develop that can extract and interpret other contextual clues.

One such approach is to focus on the detection of surrounding *entities* and use these as contextual clues. Besides the place itself, these entities may include other places, actors, objects, organizations, and events. Examples of such associated entities are landmarks, sports teams, well known figures such as politicians or celebrities, and nearby places that share a common administrative unit [22]. Intuitively, when a text mentions *Washington* along with *Redskins*, an American football team based in Washington, D.C., it is very likely that the *Washington* in the text refers to Washington, D.C., rather than another places called with the same toponym. It has been shown that such a co-occurrence model increases disambiguation quality [11, 18].

In addition to entities, implicit *thematic* information buried in the text can also provide contextual evidence to disambiguate place names. Similar to entities, some particular thematic topics are more likely to be mentioned along with a place, which is characterized by those topics. Topic modeling makes it possible to discover topics from the text and match texts with similar topics. Thus, given topics learned from a corpus of texts about candidate places and the topics discovered from the short text under consideration, computing a similarity score between topics representative for the text and for each of the candidate places can provide additional contextual clues [1]. For example, when people are talking about *Washington, DC*, political topics featuring terms such as *conservative*, *policy*, and *liberal* are more likely to be mentioned than when talking about the (small) city of *Washington, Pennsylvania*.

The core distinction between these perspectives is that mentioned entities are explicit information, while thematic information is usually implicit. Both types of information are used as clues by humans to disambiguate a place name. In this paper, we propose a novel approach which integrates *things and strings*, i.e., entity co-occurrence and topic modeling, thereby combining explicit and implicit contextual clues. **The contributions of this work are as follows:**

- We apply topic modeling to place name disambiguation, an approach that has not been taken before.
- We integrate this topic-based model with a reworked version of our previous entity-based co-occurrence model [11] and learn the appropriate weights for this integrated model.
- We compare the integrated model to three well known systems (TextRazor, DBpedia Spotlight, and Open Calais) as baselines and demonstrate that our model outperforms all of them.

2 Related Work

As an extension of named entity disambiguation, place name disambiguation can be conducted using the general approaches from named entity disambiguation. Wikipedia, as a valuable source for ground truth descriptions of named entities, has been used in a number of studies. For example, Bunescu and Pasca [5] trained a vector space model to host the contextual and categorical terms derived from Wikipedia, and employed TF-IDF to determine the importance of these terms. Milne and Witten [17] describes a method for augmenting unstructured text with links to Wikipedia articles. For ambiguous links, the authors proposed a machine learning approach and trained several models based on Wikipedia data. Two named entity disambiguation modules were introduced by Mihalcea and Csomai [16]. One measured the overlaps between context and candidate descriptions, and the other trained a supervised learning model based on manually assigned links in the Wikipedia articles.

For studies specifically focusing on place name disambiguation, Jones and Purves [13] discussed using related places to resolve place ambiguity. Machado et al. [14] proposed an ontological gazetteer which records the semantic relations between places to help disambiguate place names based on related places and alternative place names. In a similar approach, Spitz et al. [22] constructed a network of place relatedness based on English Wikipedia articles. Zhang and Gelernter [24] proposed a supervised machine learning approach to rank candidate places for ambiguous toponyms in Twitter messages that relies on the metadata of tweets and context to a limited extent. In previous work, we leveraged the structured Linked Data in DBpedia for place name disambiguation and demonstrated that a combination of Wikipedia and DBpedia data leads to generally better performance [11].

3 Methodology

The work at hand differs from these previous studies. We apply topic modeling for place name disambiguation and integrate the trained topic model with an entity-based model which captures the co-occurrence relations. Thereby we combine a *things*-based perspective with a *strings*-based perspective.

In the following, we assume that the *surface forms* of place names have been extracted prior to disambiguation, so the primary task of place name disambiguation is to identify the place to which a surface form refers. To accomplish this a list of candidate entities, i.e., places, is selected. In prior work, knowledge bases, such as Wikipedia, DBpedia, and WordNet have been used to obtain candidate entities [6, 15, 10], and here we employ DBpedia as the source of candidate entities. Once a set of candidate places has been identified, the likelihood that the surface form refers to each entity is measured and the disambiguation result is returned if the computed score exceeds a given threshold.

3.1 Entity-based Co-Occurrence Model

In this section we describe the entity-based co-occurrence method. Wikipedia and DBpedia are used as the sources to train our model. We define the entities from Wikipedia as those words or phrases on a Wikipedia page of the candidate places which have links to another page about these entities. The entities from DBpedia are either subjects or objects of those RDF triples which contain the candidate place entities. Not all RDF triples are selected, but those that fall under the DBpedia namespace, i.e., with prefix *dbp*[6] and *dbo*.[7] While *dbo* provides a cleaner and better structured mapping-based dataset, it does not provide a complete coverage of the original properties and types from the Wikipedia infoboxes. In order to avoid data bias we use both *dbo* and *dbp*. Literals were excluded as well. We treat the subject or object of a triple as a whole, i.e., as an individual entity, instead of further tokenizing it into terms. The harvested entities differ greatly. They include related places (of different types), time zone information, known figures that were born or died at the given place, events that took place there, companies, organizations,[8] sports teams, as well as representative landmarks such as buildings or other physical objects.

Table 1 shows some sample entities for Washington, Louisiana, derived from Wikipedia and DBpedia. It should be noted that there is considerable overlap between place data extracted from Wikipedia and DBpedia. Moreover, some properties such as *population density* in Wikipedia can occur for most or even all candidate places. Such entities which appear frequently but help less to uniquely identify a place will not play a crucial rule in disambiguating the place names.

The entities are assigned weights according to their relative connectivity to the places by means of *term frequency-inverse document frequency (TF-IDF)*. The term frequency of the entity is the number of times the entity appears in Wikipedia and DBpedia, so in this case, it could be 0, 1, and 2. We only count each entity's appearance in a document once, so the term frequency will not be inflated by those entities which are related to many candidate place entities while contribute less to uniquely identify the place. The formula of applying TF-IDF to assign weights to entities is defined in Eq. 1, 2, and 3.

[6] http://dbpedia.org/resource/
[7] http://dbpedia.org/ontology/
[8] For example via dbr:FreedomWorks dbp:headquarters dbr:Washington,_D.C. .

Washington, Louisiana
Wikipedia — St.Landry Parish; Opelousas; Eunice; population density; medianhousehold income; American Civil War; Connecticut; cattle; cow; corn...
DBpedia — United States; Central Time Zone; St. Landry Parish, Louisiana; John M. Parker; KNEX-FM; Louisiana Highway 10...

Table 1: Sample entities for Washington, LA, from Wikipedia and DBpedia

$$tf(e) = \begin{cases} 0 & e \text{ is not in Wikipedia and DBpedia} \\ 1 & e \text{ is either in Wikipedia or DBpedia} \\ 2 & e \text{ is in both Wikipedia and DBpedia} \end{cases} \tag{1}$$

$$idf(e) = 1 + log(\frac{|E|+1}{n_e}) \tag{2}$$

$$Weight(e) = ID\text{-}ITF(e) = tf(e) \times idf(e) \tag{3}$$

Here $tf(e)$ defines the term frequency of an entity e, and $idf(e)$ defines the inverse document frequency of e. $|E|$ is the number of all potential candidate places for a surface form, and n_e represents the number of candidate places which contain the entity e. Using TF-IDF entities appearing in multiple candidate places are given lower weights, while entities which are able to uniquely identify a place have more weights. For example, the fact that a place is within the United States becomes irrelevant as it holds for all of them.

We then measure the likelihood that a surface form in a test sentence refers to a candidate place through an entity matching score. To calculate the entity matching score, we first find those entities of the candidate place which also appear in the short text. The weights of matching entities are summed to produce an entity matching score of the candidate place to the surface form in the test sentence. The score is calculated as given in Eq. 4.

$$S_{EC}(s \rightarrow c_i) = \sum_{j=1}^{m} (Weight(e_j) \times I_j) \tag{4}$$

Here m corresponds to the number of entities e for the candidate c_i. I_j is either 1 or 0, referring to whether a matching entity is found in the test for the entity e_j. The candidate place with higher entity matching score is regarded to more likely be the actual place to which the surface form refers. The matching score is the final output of the entity co-occurrence model.

3.2 Topic-based Model

In this section we introduce the topic-based model. It makes use of the fact that text is *geo-indicative* [1] even without having any direct geographic references. Hence, even everyday language should be able to provide additional evidence

for place name disambiguation. For example, terms such as *humid*, *hot*, *festival*, *poverty*, and even *American Civil War* are more likely to be uttered when referring to Washington, Louisiana than Washington, Maine. The latter rarely experiences hot and humid weather, does not host a popular festival, has substantially less poverty problems compared to its namesake, and did not play a notable role in the civil war. Here we use Latent Dirichlet allocation (LDA) for topic modeling. LDA is a popular unsupervised machine learning algorithm used to discover topics in a large document collection [4]. Each document is modeled as a probability vector over a set of topics, providing a dimensionally-reduced representation of the documents in the corpus.

We use the geo-referenced text from the English Wikipedia as the source material for discovering these thematic patterns. We start with the idea that a collection of texts that describe various features in a local region–such as museums, parks, mountains, architectural landmarks, etc.–give us a foundation for differentiating places referenced in other texts based on thematic, non-geographically specific, terms. For this we need a systematic way to associate the training documents in Wikipedia with well-defined regions. Because administrative regions vary widely in area, they do not provide a good mechanism for aggregation. Instead, our solution is to aggregate the geo-referenced texts in Wikipedia based on an equal area grid over the Earth. This solution means that articles with point-based geo-references are binned together if they spatially intersect with a grid cell, while text related to areal features (such as national parks) can be associated with multiple grid cells.

There are several options for creating a discrete global grid based on an polyhedral simplification of the Earth [21]. In this work we utilize the Fuller icosahedral Dymaxion projection to create a hierarchical triangular mesh [9]. The triangular mesh can be made successively more fine-grained by dividing each triangle into four internal triangles. For place name disambiguation we need grid cells that are fine-grained enough so that two possible places with the same name do not fall within one grid cell. The Fuller projection at hierarchical level 7 (shown in Figure 2) provides a mesh over the Earth with 327,680 cells with inter-cell distance of 31.81 km and cell area of 1,556.6 km^2, sufficient to handle most place name disambiguation tasks for meso-scale features like cities.

Once we identified all articles that have geo-references that spatially intersect with a grid cell we can combine all the text to create a *grid document*. For the English Wikipedia the geo-referenced articles intersect with 63,473 grid cells at Fuller level 7. The resulting 63,473 grid documents serve as the training data input for LDA topic modeling. We utilized the MALLET implementation of LDA with hyperparameter optimization, which allows for topics to vary in importance in the generated corpus, and we trained the LDA topic model with 512 topics.

The MALLET toolkit generates an inferencer file for testing new documents against a trained LDA model. For a new document or snippet of text, we use the trained topic model to infer the most likely candidate location based on the inferred mixture of topics. Given a set of candidate locations (i.e., point coordinates) we find the topic mixtures for the grid cells that spatially intersect the

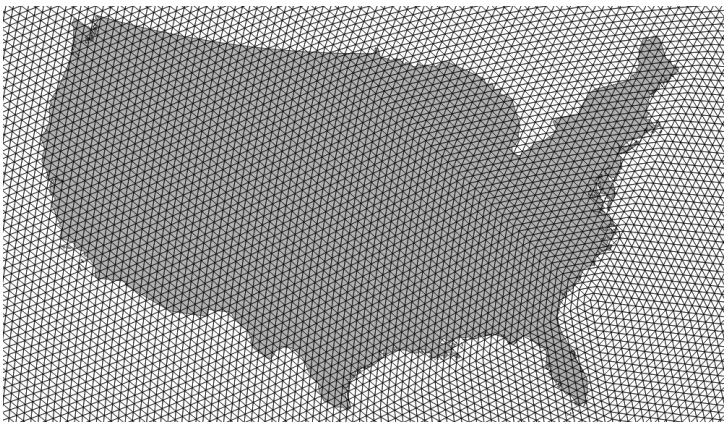

Fig. 2: Level 7 triangular mesh discrete global grid built using Fuller icosahedral Dymaxion projection, shown in U.S. Contiguous Albers projection.

locations and calculate the Jensen-Shannon divergence (Eq.6) between probability vector representations of the topic mixtures for each candidate and the topic mixture for the new document. The JS divergence is a symmetric measure calculated from the average of the relative entropies (Kullback Leibler divergence, shown in Eq. 5) between two probability vectors (P and Q) and their average, $M = \frac{1}{2}(P + Q)$. The JS divergence is a standard measure of similarity between two probability vectors, and is commonly used for calculating similarity based on topic model results [23]. A lower JS divergence result indicates greater thematic similarity between the new text and the candidate location.

$$KL(P \parallel Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)} \tag{5}$$

$$JS(P \parallel Q) = \frac{1}{2} KL(P \parallel M) + \frac{1}{2} KL(Q \parallel M) \tag{6}$$

3.3 Integrated Model (ETM)

The first model makes use of the co-occurrence of entities as contextual clue to disambiguate place names, while the second model puts emphasis on linguistic aspects, namely co-occurring topics. As argued in the introduction, applying a single model, which extracts partial contextual clues, is often not sufficient to differentiate place names from short texts. Thus, we combine the entity-based model and string-based topic model to an integrated approach called ETM (Entity & Topic Model).

Both the entity co-occurrence model and the topic-based model return a score when comparing each candidate place with each ambiguous place name in a sample text. The scores from these two models are not directly comparable as they involve relative probabilistic measures. To combine the models, we must

first standardize the scores of the candidate places for each short text. This results in setting the standardized mean to zero. Scores originally higher than the mean will be positive, and scores originally lower than the mean will be negative. For each candidate place name, the standardized scores from the entity co-occurrence model are then combined with the standardized scores from the topic-based model along with a weighting parameter λ as shown in Eq. 7.

$$S_{ETM}(s \to c_i) = \lambda S_{ECM}(s \to c_i) + (1 - \lambda) S_{TM}(s \to c_i) \tag{7}$$

Here $\lambda \in [0, 1]$, and determines how much each model is weighted in the combined approach. S_{ECM} is the standardized score computed from the entity co-occurrence model for the candidate place name c_i with respect to the surface form s, while S_{TM} is the standardized score from the topic model, namely the JS divergence. S_{ETM} is the score of the combined model, which is the sum of the weighted standardized scores of the two models. Provided that S_{ETM} is the probability of a candidate place which a surface form refers to, the percentile is used as the threshold over which candidate places are returned as the disambiguation result.

4 Evaluation

In this section we evaluate the performance of our proposed ETM and describe the methods through which we gathered the testing corpus and the metrics employed for the evaluation.

4.1 Preparing the Test Corpus

We constructed a text corpus specifically for the evaluation of our place name disambiguation models. The corpus is used to evaluate the performance of the combined ETM and to compare it to existing systems acting as baselines.

Oxford, Wisconsin — Located in Marquette County in south-central ~~Wisconsin~~, just minutes west of Interstate 39, Oxford invites you to experience our small town charm along with the area's many year-round outdoor attractions.
Jackson, Montana — The tiny town of Jackson, ~~Montana~~ has made a name for itself as a winter sports destination for the adventurous.
Dayton, Nevada — Since the Native-American tribes in the area were nomadic, this made Dayton the first and oldest permanent non-native settlement in ~~Nevada~~.

Table 2: Three example records of the test corpus extracted from websites.

To construct the corpus, we first derive ambiguous place names from a list of the most common U.S. place names on Wikipedia.[9] As the list also presents

[9] https://en.wikipedia.org/wiki/List_of_the_most_common_ U.S._place_names

the full place names which could be used to identify the place of interest, we feed the full place names into the Bing Search API,[10] which returns a list of websites related to the place along with URLs. URLs containing "Wikipedia" are filtered out. We then visit the selected websites and extract sentences which contain the full place name. The auxiliary part of the full place name (state or county name) is removed, so the remaining place name is ambiguous. The result of this approach is a set of real-world, i.e., not synthetic, sentences containing ambiguous place names. These sentences comprise our ground truth data.

Sample ground truth sentences are shown in Table 2. The full place name and test sentence are separated by an em-dash, and the auxiliary part of the full place name is removed (shown as *striken* for example purposes). This resulting data contains noise. Some sentences, for instance, contain no meaningful entities or terms that can be categorized into topics, while others seem to be automatically generated from templates. This noise, however, can help evaluate the robustness of our models. In total, the testing corpus consists of 5,500 sentences. The average length of a test sentence is 22.54 words with a median of 19. Note that stop words count towards these statistics, while auxiliary parts of the place name do not.

4.2 Metrics

F-score and Mean Reciprocal Rank (MRR) are used as metrics for the performance evaluation of the place name disambiguation models. The F-score (see Eq. 8) is defined as the harmonic mean of precision and recall [3]. MRR, by comparison, considers the order of the results; see Eq. 9. The reciprocal rank of a test sentence is the inverse of the rank of the correctly identified place name in the list of the candidate places for the surface form.

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{8}$$

$$MRR = \frac{1}{|Q|} \sum_{r=1}^{|Q|} \frac{1}{rank_i} \tag{9}$$

4.3 Results

In this section, we present the results of our evaluation and compare them to other well recognized named entity disambiguation systems as baselines.

DBpedia Spotlight[11], TextRazor[12], and Open Calais[13] were selected as baseline systems to be compared to ETM. DBpedia Spotlight is based on DBpedia's rich knowledge base of structured data [15], which is also employed by our proposed model. Two endpoints of DBpedia Spotlight Web Service (V. 0.7) were

[10] https://datamarket.azure.com/dataset/bing/search
[11] https://github.com/dbpedia-spotlight/dbpedia-spotlight
[12] https://www.textrazor.com/
[13] http://www.opencalais.com/

used for testing, namely *Annotate* and *Candidates*. The *Candidates* endpoint returns a ranked list of candidates for each recognized entity and concept, while *Annotate* simply returns the best candidate according to the context. TextRazor and Open Calais are two commercial Web services for named entity recognition and named entity disambiguation. Both services offer application programming interfaces (APIs). The TextRazor API returns only one candidate for each entity recognized from the test sentence. Experiments were conducted [20] to compare several named entity disambiguation systems which included DBpedia Spotlight (V. 0.6, confidence=0, support=0) and TextRazor. In the experiments, TextRazor demonstrates the best performance in terms of F-score. Open Calais API also returns only one candidate for each recognized entity, while it provides additional *social tags* for each test text instance.

Given that TextRazor and Open Calais do not provide controls on how many candidate places are returned and DBpedia Spotlight relies on *Confidence* and *Support* which are not comparable to percentiles, we choose the highest scores each baseline systems can reach to compare it to our models. For instance, for DBpedia Spotlight, we picked $Confidence = 0.2$ and $Support = 0$, given that this combination of parameter leads to the best overall performance for our setting. From Figure 3 we can see that Open Calais can obtain relatively higher F-scores and MRR than TextRazor and DBpedia Spotlight on the test corpus. The F-score and MRR of those baseline systems on the testing dataset are shown in Table 3. Compared to these systems, the *individual* performance of the entity-based co-occurrence model and topic-based model do not show a significant improvement, except for the entity co-occurrence model on MRR.

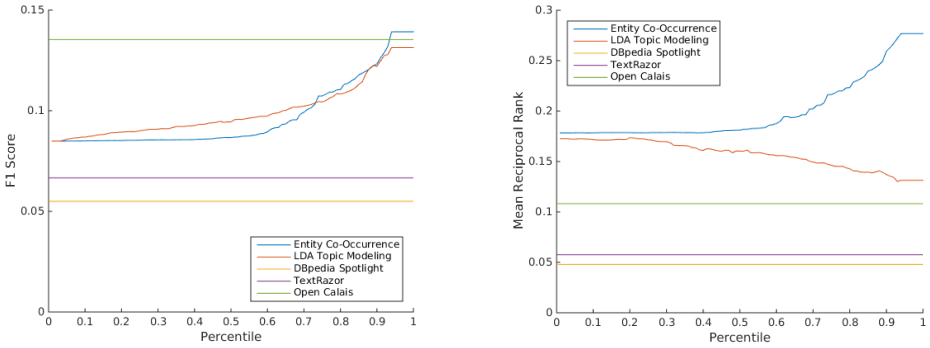

Fig. 3: (left) F-score and (right) Mean Reciprocal Rank for the entity co-occurrence model and the topic model along percentile, and comparison with DBpedia Spotlight, TextRazor, and Open Calais.

Figure 3 also shows how F-score and MRR change along percentiles. Note that the 0.9 at the x-axis refers to the 90th percentile, which means that the candidate places with top 10 percentage of scores are selected as the disam-

biguation result. As shown in the plots, when percentile increases, the F-scores of both individual models increase very slightly until the 60th percentile when the scores start increasing dramatically. The MRR for the entity co-occurrence model along percentiles has a similar trend as the F-score, while the MRR for the topic model drops when less candidate places are selected.

ETM, which combines the entity-based co-occurrence model with the topic-based model, demonstrates a significant improvement in terms of F-score and MRR, as shown in Figure 4. We tested λ values from 0 to 1 with an interval of 0.01 and found that $\lambda = 0.48$ yields the best results on the test dataset. This indicates that both the entity co-occurrence model and the topic model play roughly even roles in ETM for disambiguating place names. At the 94th percentile, the F-score is 0.239, while MRR is 0.239. Note that F-score and MRR are different values though they happen to be rounded to the same value. Out of 5,500 test sentences, 1,315 are correctly disambiguated, given the disambiguation result of 5,509 places. The figures show that both F-score and MRR increase along with percentiles and reach peaks when very low percentage of records are returned as disambiguation results.

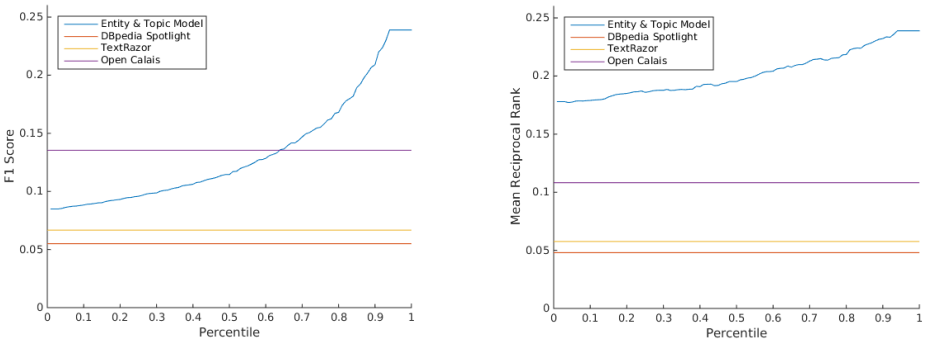

Fig. 4: (left) F-score and (right) Mean Reciprocal Rank for ETM ($\lambda = 0.48$), DBpedia Spotlight, TextRazor, and Open Calais.

In some cases, only one candidate (if available) is taken as the disambiguation results. As stated in the previous paragraph, TextRazor only outputs at most one result, so does DBpedia Spotlight Web Service in the *Annotation* mode. For Open Calais, the disambiguation result is ranked, so the first returned result is taken for this evaluation. When only the candidate places with highest scores are taken, the F-score for ETM reaches 0.238 when λ is set to 0.48. Since always one candidate is picked for each testing sentence, predicted condition positives are the same as condition positives. Thus, the mean reciprocal rank, precision and recall are identical, and they top at 0.238 with λ being 0.48. The change of F-score and Mean Reciprocal Rank for our proposed ETM along λ and its comparison to DBpedia Spotlight, TextRazor, and Open Calais are shown in

Figure 5. As shown in the figure, with the increase of λ, after the peak when λ is around 0.48, F-score and MRR drop mildly until λ approaching 1 when F-score and MRR drop significantly. This implies the entity co-occurrence model plays a more important role for this task, while the topic model still helps to improve the performance. Out of 5500 testing sentences, EMT is able to correctly identify 1311 ambiguous places.

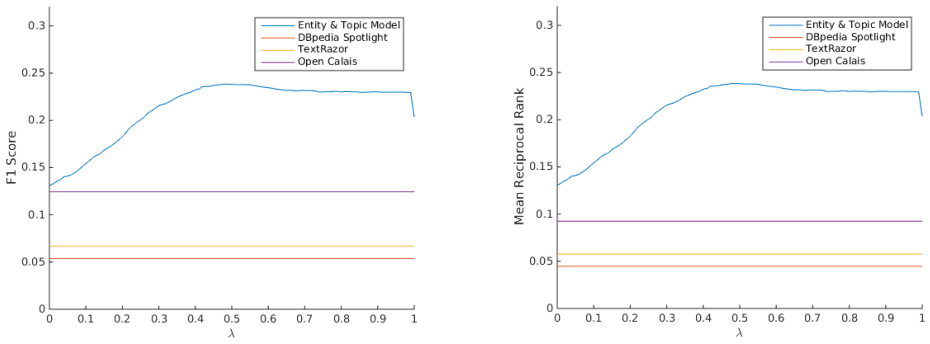

Fig. 5: (left) F-score and (right) Mean Reciprocal Rank for ETM, DBpedia Spotlight, TextRazor, and Open Calais, when only the best candidate entity is taken.

The evaluation of ETM and its comparison to baseline systems are summarized in Table 3. Overall, based on the evaluation, the proposed ETM substantially outperforms existing named entity disambiguation systems in terms of F-score and Mean Reciprocal Rank. The fact that all F-scores are low, is an important reminder for the fact that place name disambiguation from short texts is a difficult task (and that some test sentences did not contain any or only minimal contextual clues).

Model	Parameters	Precision	Recall	F1-Score	MRR
DBpedia Spotlight	$Confidence = 0.2$; $Support = 0$	0.057	0.053	0.055	0.048
TextRazor	n/a	0.070	0.063	0.067	0.058
Open Calais	n/a	0.148	0.125	0.135	0.108
ETM	$\lambda = 0.48$; 94th percentile	0.239	0.239	0.239	0.239

Table 3: Comparison of systems at best performance in terms of Precision, Recall, F1-Score and Mean Reciprocal Rank (MRR)

5 Conclusions and Further Work

In this paper we proposed a novel approach to tackle the challenging task of disambiguating place names from short texts. Place name disambiguation is an important part of knowledge extraction and a core component of geographic information retrieval systems. We have presented two models that are driven by

different perspectives, namely an entity-based co-occurrence model and a topic-based model. The first model focuses on the semantic connections between entities and thereby on *things*, while the second model works on the linguistic level by investigating topics associated with places and thereby takes a *string*-based perspective. The integration of both models (called ETM) shows a substantially better performance than the used baseline systems with respect to F-score and MRR.

Nonetheless, there is space for future improvements. For the entity-based model, properties other than those with namespaces of *dbo* and *dbp* have been filtered out. The same is true for literals. Both of these could be added to a future version of ETM, although they would require more work on the used similarity functions in case of the literals and a better alignment to ensure that properties from different namespaces are not mere duplicates. In our work, the ETM is realized as a convex combination of the entity-based co-occurrence model and the topic-based model. Other approaches could be investigated as well. We have used LDA for topic modeling but this is not the only choice that can be used and other approaches will be tested in the future.

As for the experiment, although place entities in our testing corpus have highly ambiguous place names, those places are all some kind of administrative divisions (i.e., cities, towns, villages, etc.) and located within the United States. A potential improvement could be seeking more ambiguous place names from other types of places which are outside of the United States.

Acknowledgement

The authors would like to acknowledge partial support by the National Science Foundation (NSF) under award 1440202 EarthCube Building Blocks: Collaborative Proposal: GeoLink Leveraging Semantics and Linked Data for Data Sharing and Discovery in the Geosciences.

References

1. B. Adams and K. Janowicz. On the geo-indicativeness of non-georeferenced text. In *International AAAI Conference on Web and Social Media (ICWSM)*, pages 375–378, 2012.
2. B. Adams, G. McKenzie, and M. Gahegan. Frankenplace: interactive thematic mapping for ad hoc exploratory search. In *Proceedings of the 24th International Conference on World Wide Web*, pages 12–22. ACM, 2015.
3. M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23, 2003.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.
5. R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, volume 6, pages 9–16, 2006.
6. S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716, 2007.

7. A. Fader, S. Soderland, O. Etzioni, and T. Center. Scaling wikipedia-based named entity disambiguation to arbitrary web text. In *Proceedings of the IJCAI Workshop on User-contributed Knowledge and Artificial Intelligence: An Evolving Synergy, Pasadena, CA, USA*, pages 21–26, 2009.
8. M. F. Goodchild and J. A. Glennon. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3):231–241, 2010.
9. R. W. Gray. Exact transformation equations for Fuller's world map. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 32(3):17–25, 1995.
10. X. Han and J. Zhao. Structural semantic relatedness: a knowledge-based method to named entity disambiguation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 50–59. Association for Computational Linguistics, 2010.
11. Y. Hu, K. Janowicz, and S. Prasad. Improving wikipedia-based place name disambiguation in short texts using structured data from dbpedia. In *Proceedings of the 8th Workshop on Geographic Information Retrieval*, page 8. ACM, 2014.
12. K. Janowicz and P. Hitzler. The digital earth as knowledge engine. *Semantic Web*, 3(3):213–221, 2012.
13. C. B. Jones and R. S. Purves. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228, 2008.
14. I. M. R. Machado, R. O. de Alencar, R. de Oliveira Campos Jr, and C. A. Davis Jr. An ontological gazetteer and its application for place name disambiguation in text. *Journal of the Brazilian Computer Society*, 17(4):267–279, 2011.
15. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8. ACM, 2011.
16. R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM, 2007.
17. D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.
18. S. Overell and S. Rüger. Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, 22(3):265–287, 2008.
19. R. Purves and C. Jones. Geographic information retrieval. *SIGSPATIAL Special*, 3(2):2–4, 2011.
20. G. Rizzo, M. van Erp, and R. Troncy. Benchmarking the extraction and disambiguation of named entities on the semantic web. In *LREC*, pages 4593–4600, 2014.
21. K. Sahr, D. White, and A. J. Kimerling. Geodesic discrete global grid systems. *Cartography and Geographic Information Science*, 30(2):121–134, 2003.
22. A. Spitz, J. Geiß, and M. Gertz. So far away and yet so close: Augmenting toponym disambiguation and similarity with text-based networks. In *Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data*, GeoRich '16, pages 2:1–2:6, New York, NY, USA, 2016. ACM.
23. M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.
24. W. Zhang and J. Gelernter. Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, 2014(9):37–70, 2014.