Multimedia Systems (2016) 22:443–453
DOI 10.1007/s00530-014-0403-z

SPECIAL ISSUE PAPER

Resolving local cuisines for tourists with multi-source social media contents

Zhao-Yan Ming · Tat-Seng Chua

Published online: 24 July 2014

Abstract Social media are playing an increasingly important role as information sources for tourists. However, the research on increasing the accessibility of social media for tourists is sparse. In this work, we study local cuisine as a specific application of the problem. At the social media side, food has become one of the highest rated topics with the popularity of mobile devices. Texts, photos, and videos about various cuisines are produced when people tweet, blog, and interact with friends in the social networks in their daily life. On the other hand, tourists often encounter cuisine names that they have difficulties in understanding the meanings, even with the help of a dictionary. Comprehensive heterogenous social media contents about local cuisines thus can greatly help in filling the gap. In this work, we propose utilizing the multi-source social media contents to resolve the local cuisines for tourists. The overall approach consists of two major components. First, a location-aware linking module is built to resolve the cuisine names, especially the pseudonym issue where the same food is known differently in various contexts. Second, given the resolved cuisine names, a location-aware aggregation module is designed to compile the relevant social media contents in various media forms. Experiments demonstrate the effectiveness of the two modules. Furthermore, a user study shows that both the linking module and the aggregation module are helpful for tourists entering localities of new food.

Keywords Multi-source social media content · Location-aware aggregation · Local cuisines

1 Introduction

Social media are playing an increasingly important role as information sources for tourists. However, the research on increasing the accessibility of social media for tourists is sparse. In work, we study local cuisine as a specific application of the problem. Often, tourists in a foreign country encounter dish names that they have no idea of or even a wrong idea about what is in them, even aided with a dictionary. For example, people may roughly know from the names that "Caesar salad" is a salad and "lobster Newburg" is made of lobster, but they may not know how they are different from the ordinary salad and lobster. It is desirable that the local cuisines are presented in an enriched form rather than with enriched names for the tourists to make better dining decisions and even learn the recipes and stories behind them.

On the other hand, food is becoming one of the highest rated topics in social media. It is popular for people to text, take photos or socialize on a mobile device at mealtime, as a means of checking in with peers on social media sites like Twitter, Facebook and personal blogs. Moreover, local chefs and food bloggers also like to tweet recipes, post how-tos and make videos of their food creations. These timely, personal, relevant social media contents about local

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

Z.-Y. Ming (✉)
National University of Singapore, AS6 0419, Singapore, Singapore
e-mail: mingzhaoyan@gmail.com

T.-S. Chua
National University of Singapore, AS6 0508, Singapore, Singapore

cuisines can be an interesting and useful source to enhance the dining experiences of tourists.

To link these online activities by the local people with the real-world need of tourists, we thus propose a novel task of resolving local cuisines with social media contents. The task can be solved in two steps: first, the ambiguity of the cuisine name is resolved within the context by linking with the knowledge base entries. Second, given the clearly defined cuisine name in the knowledge base, relevant and high-quality multi-source social media contents are gathered to provide comprehensive user-centric information.

For the first step, research efforts have been devoted to linking general entities [19, 25, 35], such as persons, locations and organizations with their entries in the KB. However, there are few works on entities from specific domains. Though similar in nature, the general entity-linking task and the cuisine name linking task have different challenges. In entity linking, the ambiguity of entities posits a major difficulty. However, the challenge of resolving of cuisine names is mainly due to the fact that a dish may have multiple names, called the pseudonym issue. For example, the Cantonese noodle dish “Wonton noodles” which is popular in Guangzhou, Hong Kong, Singapore, Thailand and Malaysia, may be called as “Yun-tun mian”, “Wan-tan Min” and “wanton mee” depending on the location it is served. Ambiguity in dish names such as “cockle” is less of a problem as long as we only consider entities that are edible rather than the entity “cockle” as a canon name. Therefore, existing works that usually focus on resolving the polysemy (ambiguity) issue [19, 25, 35] are not readily applicable to the dish name linking task.

For the second step, the social media contents from multiple sources are gathered, as information in any single source is always limited and domain specific. For example, if a user wants to know the general description for “Wonton noodles”, it is better to refer to Wikipedia. When he wants to know how the local people like “Wonton noodles”, blogs and twitter turn out to be better places. If he enjoys the noodle and wants to make it himself at home, a relevant YouTube video can show him the recipes and steps. As a result, to present an overall picture of the cuisine, the system is required to integrate these heterogeneous contents with an emphasis on diversity. Moreover, the inherent uneven quality of the contents can be an alarming issue for proper user experiences. Though pioneer work on social media content integration [37] has successfully mined the different characteristics of multiple sources, they usually ignore the usability issue such as the content quality.

In this work, we propose targeting modules for the above-discussed problems. We first propose a location-aware cuisine entity-linking algorithm to find for the target cuisine its name variants and associated knowledge base description. Specifically, a geolocator is proposed to utilize meta information and mine the intrinsic content. The geocontext is further used in the entity-linking model for retrieving the name variants and enhancing the accuracy of the cuisine name resolution. We then use the geocontext, the name variants and the knowledge base definitions to locate relevant social media contents from multiple sources. In particular, we propose a content quality estimator that works seamlessly with a diversity-focused relevance ranking model. We empirically evaluate our proposed method for resolving cuisines on a food review dataset with a good mixture of Chinese, Malay and Indian dishes. Experiments demonstrate the effectiveness of the two modules. Furthermore, a user study shows that both the linking module and the aggregation module are helpful for tourists entering localities of new food.

To tackle the problems in the cuisine resolution task, we introduce a dedicated location-aware entity-linking model that focuses on the pseudonym issue. The contributions of this work are twofold:

- First, we propose a novel cuisine name resolution task that links the real-world needs of tourists with the online social media contents generated by the local people.
- Second, we propose a location-aware entity-linking model for cuisine name resolution, and a multi-source social media contents aggregating model that emphasizes on the diversity and quality of contents in addition to the location-enhanced relevance.

The remaining sections are organized as follows. We first review related literatures in Sect. 2. Section 3 presents an entity-linking approach to resolving cuisine names using knowledge bases. Section 4 details the method for enriching cuisine knowledge with heterogenous social media contents. We conduct empirical evaluations in Sect. 5 and conclude the work in Sect. 6.

2 Related work

2.1 Entity linking with knowledge bases

Though there has been no previous work linking dish names to KB to our knowledge, works on providing semantic links for various contents are not new. For example, Odijk et al. [29] proposed relevant links to Wikipedia articles for TED talk based on the subtitles. Generally referred to as entity linking, the task has been undertaken in the pioneering work of in [5, 10, 27, 28]. Bunescu and Pasca [5] proposed linking persons' names in Washington Post news articles with their referent Wikipedia article, while [10] proposed linking all entities mentioned in news articles. Mihalcea and Csomai [27] developed Wikify, a system which automatically generates links within Wikipedia by first detecting and then identifying the terms from which links should be made. Following them, a task on entity linking has been organized under the Knowledge Base Population track at TAC since 2009 [25], where person, organization and geopolitical entities appearing in news articles are linked with Wikipedia entries. Entity linking with KB is also closely related to cross-document entity coreference, a process of determining whether or not two mentions of entities in different documents refer to the same entity [4, 12, 22, 24].

Both unsupervised and supervised approaches have been proposed. In the unsupervised case, [10] proposed a similarity-based vector space model for linking an ambiguous mention in a document with an entity in Wikipedia. Han and Zhao [16] proposed a generative probabilistic model to leverage heterogenous entity knowledge (including the entity popularity, name, and context) for the entity-linking task. Varma et al. [32] adopted an information retrieval approach where they indexed the KB and handled queries that are acronym and not acronym separately. More recently, [15] proposed linking entity based on a statistical language model-based information retrieval model with query expansion.

In contrast to the unsupervised approaches, supervised methods have the advantage of combining the features without the need to tune the rules or weights for the specific datasets. For the selection of learning models, the entity-linking task of one query with multiple candidates is cast into either a classification model [2, 5, 28, 35] or a ranking model [11, 36]. Classification models are learned from labeled data to determine whether or not a query refers to the candidate entity. For supervised ranking model, a list of candidate entities is considered for a given query. Learning-to-rank models such as SVM_Rank [6, 21] and ListNet [7] are employed to rank the candidates so as to pick out the best one. While [11]'s ranking model is a pairwise approach, [36] adopted the listwise methods where the constraints are based on a list of candidates for a query in the training set. Zheng et al. [36] compared learning-to-rank methods with two classification methods and found that the learning-to-rank methods were significantly better than the classification methods. However, these methods usually use a large number of features targeting at the ambiguity issue; fewer works consider the pseudonym issue or both.

2.2 Location-based social media analysis

Cheng et al. [9] proposed and evaluated a probabilistic framework for estimating a Twitter user's city-level location based purely on the content of the user's tweets, even in the absence of any other geospatial cues. Amitay et al. [3] described Web-a-Where, a system for associating geography with Web pages. Web-a-Where locates mentions of places and determines the place each name refers to. In addition, it assigns to each page a geographic focus—a locality that the page discusses as a whole. To solve the inherent ambiguity of location relevance, [34] performed location relevance mining in two complementary angles, relevance classification and relevance ranking, to provide comprehensive understanding of locations. Jiang et al. [20] presented a local specialty mining algorithm, which utilized both the structured data from local review websites and the unstructured user-generated content from community Q&A websites and travelogues. The proposed algorithm extracted dish names from local review data to build a document for each city and applied tfidf weighting algorithm on these documents to rank dishes.

2.3 Multi-source integration

Information from multiple sources provides researchers clues of different views and helps to achieve better results by overcoming the bias of any single information source [13, 14, 17, 18, 37]. Han et al. [17] used information from Wikipedia, WordNet and an NE co-occurrence corpus to measure the semantic relatedness between words. Although they just chose the conditionally most confident source to estimate the relatedness, instead of integrating the three sources together, the results have already outperformed the methods which only used a single source. On the other hand, Hoffart et al. [18] proposed to integrate information from Wikipedia, WordNet, Geo-Name corpus, etc., to build Yago2, an open domain structured knowledge base.

3 Resolving cuisine names with knowledge base

To understand a cuisine name, the first step is to link it with some standard knowledge, for example, the description in a knowledge base. Therefore, we start the cuisine resolution task by linking the cuisine name in any text message with a matching entry in the knowledge base.

3.1 Cuisine entity-linking problem

The cuisine name linking task has two inputs: the query containing at least one dish name and the knowledge base of all possible dishes available. Formally, the query can be denoted as $Q : (q_m, q_c)$, where q_m denotes the dish name in the query and q_c denotes the q_m's context document. A KB $\mathcal{K} = \{E_i : i = 1, \ldots, N\}$ comprises N entities where each entity $E : (e_m, e_d)$ is represented by its name e_m and description e_d. As the size of $\mathcal{K}$ is usually in millions, to reduce the computational cost, candidate generation is usually performed as a first step. A subset $\mathcal{K}'$ of $\mathcal{K}$ with a size of M $(M \ll N)$ entities are actually examined. To account for cases when a query finds no matching entry in KB, a NIL entity is added for uniform processing [26]. Therefore, the actual working KB $\mathcal{K}'$ includes the NIL entity plus the generated candidates for the query.

With the two inputs, the dish name linking task can be formulated as a ranking problem that identifies from $\mathcal{K}'$ a KB entry E^* that is equivalent to the query Q:

$$E_* = \arg \max_{E_j \in \{\mathcal{K}', NIL\}} h(Q, E_j), \tag{1}$$

where $h(Q, E_j)$ is a ranking function producing the matching score between the query Q and the candidate entity E_j. To compute the ranking score $h(Q, E)$ in Eq. 1, we take a structured approach with four matchings, they are: $f_{mm}(q_m, e_m)$ where the query dish name is matched with the KB entry name, $f_{cc}(q_c, e_c)$ where the query context is matched with the KB entry context, and $f_{cm}(q_m, e_c)$ and $f_{cm}(e_m, q_c)$, where the names are searched against one another's contexts, respectively.

Under this framework, the pseudonym is solved by modeling name variants on both the dish name query q_m and the KB title e_m using entity's coreferents and external world knowledge, to expand them to be two sets for a potential match. Polysemy of a name mention is disambiguated by matching the query context q_c and the KB description e_c. These are translated into the feature representation in the ListNet model.

Learning features: As given in the problem definition, the relation between a query dish name and a candidate entity in the KB is captured by the structured match of their names and contexts. Three categories of features are proposed, the *Name-v.s.-Name* which instantiates $f_{mm}(q_m, e_m)$, the *Context-v.s.-Context* which instantiates $f_{cc}(q_c, e_c)$, and the *Name in Context* which instantiates $f_{cm}(e_m, q_c)$ and $f_{cm}(q_m, e_c)$.

In the *Name-v.s.-Name* category, we first have the "exact match" feature. It is set to one when the query name and the entity title are exactly the same. We treat the name and its known aliases as the expanded query name. The aliases for the KB entities are extracted from the mapping lists.[1] The aliases for the query mentions are extracted from their in-document co-reference NEs. With the expanded query name set and the entity name set, we have another feature which is set to one if one of the names from either set is the same. To account for minor differences such as spelling variation, we measure the string similarity between the query name and the entity title in terms of longest common subsequence and edit distance. Similarly, as the names are expanded into sets, we take the highest string similarities between the two sets as additional features.

In the *Context-v.s.-Context* category, we have two textual features based on vector space model. The first feature measures the cosine similarity between the TF-IDF vectors of the dish review and the candidate entity article. The second feature is the cosine similarity between the TF-IDF vectors of the sentences containing the dish name and the sentences containing the entity and its co-referents.

A remaining category is the set of *Name-in-Context* features, which instantiate $f_{cm}(q_m, e_c)$ and $f_{cm}(e_m, q_c)$. Given a query q and a candidate entity e, we first consider the presence of the query name q_m and its expanded set $\mathbf{q_m}$ in the entity's context e_c. This results in two features: the number of times the query name is found in the candidate entity article; and the number of times the expanded query names appear in the candidate entity article. In the same way, we further consider the presence of the candidate entity name e_m and its expanded set $\mathbf{e_m}$ in the query's context q_c, which result in another two features.

Finally, for the NIL queries, we add some features similar to that done in [26]. We also include in the feature set the raw score of the entity passed from the candidate generation step.

4 Enrich with heterogenous social media contents

In the context of travel-related searches, social media constitute a substantial part of the search result [33]. In this work, we propose to automatically aggregate the relevant, high-quality and diverse social media content, which gives comprehensive knowledge to the tourists without the tedious manual search and selection from search results.

With cuisine name ambiguity and pseudonym resolved by the knowledge base, the targeted cuisines are now ready to be enriched with more heterogenous social media contents. Given the user-contributed nature, three problems to be solved here are the relevance, the quality, and the diversity of the contents. In the following, we will describe our proposed methods for the three problems one by one.

4.1 Geocontext enhanced relevance model

Geocontext is important for modeling cuisine information on the Web. The information on where a dish originates can be used to resolve the ambiguity and also help to identify the relevant social media contents. In this section, we propose to model cuisine geocontext in the social media content, so that only the geo-aligned contents are presented to the tourists.

Given that some social media contents are tagged with geolabel by the authors, those without tags pose a great

[1] Freebase provided Wikipedia redirects at: http://download.freebase.com/wex/ and Wikipedia hyperlink anchor-entity mapping.

challenge as there is only limited information we can get from the textual context of a cuisine name. Inspired by [9], we propose exploring the social media content posted by the same author, and use the aggregated content to identify the geolocation of the author. The identified location is then used as the geocontext of the author's social posts if they are not explicitly tagged with any location.

By aggregating across all words in tweets posted by a particular user, however, our intuition is that the location of the user will become clear.

Given the set of words $W_{sm}(u)$ extracted from a user's social media posts, we propose to estimate the probability of the user being located in city c_i as:

$$p(c_i|W_{sm}(u)) = \sum_{w \in W_{sm}(u)} p(c_i|w) \times p(w), \tag{2}$$

where $p(w)$ denotes the probability of the word w in the whole dataset. Let $ct(w)$ be the number of occurrences of the word w, and t be the total number of tokens in the corpus, we replace $p(w)$ with $count(w)/t$ in calculating the value of $p(w)$. Such an approach will produce a per-user city probability across all cities. The city with the highest probability can be taken as the user's estimated location.

$p(c_i|w)$ is the language model for the city. To estimate this parameter, we smooth the term distribution estimates for the location models using Dirichlet smoothing.

$$p(c_i|w) = \frac{p(w|c_i)p(c_i)}{p(w)}, \quad p(w|c_i) = \frac{c(w, c_i) + \mu p(t|C)}{|c_i| + \mu}, \tag{3}$$

where μ is a parameter, set empirically, $c(w, c_i)$ is the term frequency of a term w for a location c_i, $|c_i|$ is the number of terms in the location c_i, and $p(t|C)$ is the term distribution over all locations. The location-related counts are estimated using maximum likelihood from the social media contents and are tagged with a location label.

4.2 Social media content quality evaluation

Evaluation of content quality is one of the first procedures for making use of social media content in real-world applications. In our task, a quality score can be used in addition to the relevance model to enhance the usability of the results. Similar to [1], we exploit features of social media that are intuitively correlated with quality, and then train a classifier to appropriately select and weight the features for each specific type of sources and quality definition. In particular, we model the intrinsic content quality and the content popularity (votes from other users) which reflect the interaction between users. The features are then fed into a classifier that can be trained for the quality definition for the particular social media source. The difference from the original features proposed in [1] is that our features are for general text rather than a specific type.

For the first category, the intrinsic text quality features are exploited. The text features can be applied to the text items such as tweet and blogs, as well as descriptions of videos. The ten text features are:

1. Raw length in text in words.
2. Number of nouns in text.
3. Number of verbs in text.
4. Number of stop words.
5. Number of nonstop words.
6. Punctuation density in text, computed as:

$$Q\_Punct\_Dens = \frac{number\_of\_punctuations\_in\_text}{number\_of\_characters\_in\_text}. \tag{4}$$

7. Non-ASCII characters in text, computed as:

$$Q\_NonAscii\_Dens = \frac{number\_of\_non\_ascii\_chars\_in\_text}{number\_of\_characters\_in\_text}. \tag{5}$$

8. Fog score (S_{Fog}). The readability score is computed as:

$$S_{Fog} = 0.4(average\_text\_length + \%\_of\_Hard\_Words) \tag{6}$$

where *Hard Words* is the number of words with more than two syllables of a given text.

9. Flesch score (S_{Flesch}); the score is computed as:

$$S_{Flesch} = 206.835(1.015 \times \text{ASL})(84.6 \times \text{ASW}) \tag{7}$$

where ASL is the average sentence length (number of words divided by number of sentences); and ASW is the average word length in syllables (number of syllables divided by number of words).

10. For Flesch–Kincaid score ($S_{Kincaid}$), the score is computed as:

$$S_{Kincaid} = 0.39 \times \left(\frac{total\_words}{total\_sentences}\right) + 11.8 \times \left(\frac{total\_syllables}{total\_words}\right) - 15.59. \tag{8}$$

These features come from three groups: surface statistics, noise, and readability. The first group includes the statistics of the surface semantic features (1–5). It gives the number of prominent countable terms and surface semantic features of the text. We expect that high-quality text content has a good structure and contains a reasonable number of shallow syntactic features, such as verbs and nouns. The second group gives the density of special tokens which is found in the text (6–7). We expect that high-quality text contains a considerable low proportion of punctuation or special symbols. Finally, the third group is the scores of three popular readability models [30] (8–10), which estimate the educational grade level necessary to understand a portion of text based on the number of syllables detected in the given portion of text. We expect that high-quality text content contains a considerable amount of formal words and written in a good structure.

In addition to the intrinsic quality, we also try to capture the implicit and explicit votes from other users. For the explicit votes, we use the popularity signals that are available, such as the "like" counts, the number of "views", and the number of "retweet".

For the implicit votes, we use opinion analysis as a tool to infer the content popularity by analyzing past viewer's comments. We predefine three opinion categories to indicate positive, neutral, and negative rankings. After applying stop word removal and word stemming, we use punctuation, unigrams, bigrams, and part of speech (POS) bigrams to characterize each textual comment and convert it into a feature vector. The problem becomes a short-document classification problem, where we adopt a supervised classification method to classify a new comment into one of the three opinion categories. Any kind of supervised learning methods can be adopted into the system. The implicit vote for a social media message is calculated as follows:

$$iVote_{M_i} = \frac{Pos(M_i) + \kappa Neu(M_i) - Neg(M_i)}{Pos(M_i) + Neu(M_i) + Neg(M_i)}, \quad (9)$$

where $Pos(M_i)$, $Neu(M_i)$, and $Nes(M_i)$ are, respectively, the number of positive, neutral, and negative opinion labels for message M_i. $\kappa (0 < \kappa < 1)$ is the parameter to control the influence of neutral comments. The number of neutral comments can still point to the popularity of a social media message, although not as strongly as the positive comments. Overall, if a social message has a larger opinion score than others for a certain query, this message tends be to more popular from users' point of view. In our further testing, we set our parameter κ to 0.3.

Following the approach in [31], we adopt logistic regression as the supervised learner here, as it is shown to achieve a high accuracy in predicting community question answer quality. The above features are then fed into a logistic regression learner that learns and predicts the quality score of the social media messages. We then use the score to filter out the low-quality ones in retrieved results.

4.3 Content diversification

We deal with the content diversification issue in two layers. First, we try to find high-quality relevant content from each major social media source. This will ensure that different types of media are equally considered for the enrichment of the knowledge about a certain cuisine. Therefore, more efforts need to be paid on the diversification of content within sources. For example, when the relevant tweets are gathered for a cuisine, there might be some tweets that discuss about similar topics; thus the redundant ones should be reduced.

In particular, given the expanded cuisine names and their description from knowledge base, the relevant contents are first retrieved from each source and the low-quality ones are filtered out. Carbonell and Goldstein's work on maximal marginal relevance (MMR) [8] is then employed to rerank the remaining items to produce more diversified results. MMR linearly combines the relevance and the novelty that is measured against the selected items. The formula is reproduced as:

$$\mathrm{MMR} = \arg \max_{s_i \in R \backslash S} [\lambda sim_1(s_i, q) - (1 - \lambda)) \max_{s_i \in S}] sim_2(s_i, s_j)], \quad (10)$$

where q is the query, S is the set of relevant social media messages and s_i is one of S, and $sim()$ is the similarity function.

Here, $sim_2(s_i, s_j)$ can be calculated using any term-based similarity measure such as Vector Space Model. And $sim_1(s_i, q)$ is estimated using a location information enhanced relevance measure, which is defined as follows. Given the location (usually a city) c_l of the tourist and a cuisine name within a text q she/he wants to know about, the relevance between the information need and a candidate social media message s can be formulated as follows:

$$p(s|c_l, q) \approx p(s|q) \times p(c_l|s), \quad (11)$$

where $p(s|q)$ can be estimated using the language model or any retrieval model, and $p(c_l|s)$ can be approximated by $p(c_l|W_{sm}(u))$ as in Eq. 2.

5 Experiments

5.1 Cuisine name linking performance

5.1.1 Experimental settings

As there is no publicly available dataset for evaluating dish name linking, we constructed a dataset by collecting dish reviews from Singapore. The uniqueness of Singapore cuisine is that it represents a good mixture of Chinese, Malay, Indian, and western dishes. When people in Singapore write reviews in English, the dish names in the reviews can be kept in their original languages or their English translations. In practice, we collected from Foursquare the tips (short pieces of comments for a location) that belong to the famous food centers in Singapore. After filtering out the noise ones (those with less than 2 nouns or without a dish name), 1,809 tips were kept as the *short review* set. Additionally, we have a longer review set: the *blog review* set of 671 blog posts from three famous Singapore food blogs.[2] The dish names in the reviews are pre-identified, so that we focus on the task of linking rather than recognizing entities. The KB utilized is an English version of Wikipedia obtained from DBpedia,[3] which consists of 3.77 million entries as of 1 June 2012.

Two annotators manually link the dish names with the KB independently. Generally, they read the reviews and form queries to search and browse the KB until they find a match. The inter-annotator agreement is quite high: the two annotators submit almost the same labeling, except for a few reviews that one deems no linking can be found, but the other found one through more exhaustive search and browsing. This is in accordance with our discussion in previous sections: dish names linking is more a pseudonym issue than an ambiguity issue. As summarized in Table 1, the dataset contains 2,450 reviews. Of all the review queries, two-thirds of the queries (1,633) are used for training and development, and the rest of 817 are used for testing. We adopt the micro-averaged accuracy to evaluate the performance of dish name linking. As the official metric in the entity-linking task of TAC-KBP, it is defined as the ratio of the number of correctly linked queries by the total number of test queries.

Table 1 Dataset statistics

	# training		# testing	
Queries	1,663		817	
inKB	1,156	(70 %)	654	(80 %)
NIL	507	(30 %)	163	(20 %)
Blog review	447	(27 %)	224	(27 %)
Short review	1,216	(73 %)	593	(73 %)

inKB queries have a link in the KB, *NIL* queries have no link in the KB

Table 2 The overall comparison of the baselines and a state-of-the-art method

System	All queries	inKB	NIL
Title Baseline	0.709	0.653	0.932
SVM	0.791	0.768	0.883
SVM_Rank	0.824	0.804	0.905
Zheng et al. [36]	0.836^{a}	0.817^{a}	0.912
Ours	0.886^{ab}	0.879^{ab}	0.914^{a}

a and b indicate statistical significance over the baselines SVM and SVM Rank respectively at 0.95 confidence interval using paired *t* test

5.1.2 Overall results on cuisine name linking

We compare our proposed method with three baselines and a state-of-the-art method. The *Title Baseline* predicts the queries to be linked with entities of the exact title (including the title of the redirected page). SVM and SVM_Rank are two widely adopted machine learning models. Note that NIL is added as a pseudo entity for NIL prediction. Both supervised baselines are trained and tuned on the available training data with the full set of features in Sect. 3. We also implement a state-of-the-art system by [36] which uses ListNet and the features for general entity-linking task. All the supervised models work on the same candidate set generated by the recall-boosted retrieval module. Table 2 presents the end-to-end comparison between our system and three baselines. The following observations are made:

1. The Title Baseline is a strong baseline for this task. It shows that by matching a name alone we can get more than 70 % dish names linked to the right KB entry. As our supervised learning baselines, SVM_Rank works better than SVM, suggesting that ranking is more suitable than classification for the dish name linking task. This is in line with our earlier discussion that entity linking is a ranking problem given that one query is associated with multiple candidates. Of all the comparing systems, our approach achieves the highest accuracy for all queries and inKB queries with significant advantage over both supervised baselines.
2. Among the supervised ranking models, we can see that our method with the ListNet model outperforms the SVM_Rank model. Though both are learning-to-rank models, ListNet works on lists of training instances (listwise) and SVM_Rank works on pairs of instances. We conjecture that the listwise model performs better because it learns from the differences among all the candidates for a query: the resulting model captures the relations between all the candidates, while the pairwise model learns from two candidates at a time.
3. Both using ListNet, our method achieves a significantly better performance than [36]. It shows that the name modeling features work better for the dish name linking task than those features for general entity-linking task. Moreover, Zheng et al.'s [36] performance on the dish name linking data is lower than that on the official TAC data as reported in their paper. It suggests that the issues in the two data are different: the first is more on pseudonym, and the second more on ambiguity.

[2] http://ieatishootipost.sg/, http://www.ladyironchef.com/, and http://www.insing.com/.

[3] http://wiki.dbpedia.org/Downloads38.

5.1.3 *Ablation study on feature effectiveness*

From the ablation study results in Table 3, we can see that removing any group of features leads to degraded performance to some degree, suggesting that the features we designed are complementary to each other. Though all features contribute to the whole system, the different degree of degradation caused by removing the features indicates that: name-matching features are the most important group of features, followed by the context-matching features, and the name-in-context features are the least influential. This indicates that dishes are generally identifiable by their names (including the variations of names), and a relatively smaller set of dishes require context matching for disambiguation. We conjecture that name-in-context features are dependent on the length of the dish reviews and the amount of content in the candidate entity articles, and thus cannot provide constant information about query-entity relevance.

Within Context-in-Context category, we can see that removing geocontext causes substantial degradation as compared to removing all Context-in-Context features. This indicates that Geocontext is one of key context information that needs to be captured for resolving cuisine name.

5.2 Social media content aggregation

To aggregate social media content for the cuisines, we evaluate the proposed methods on content quality evaluation and geo-enhanced relevance in the following two subsections.

5.2.1 *On quality evaluation*

We now complete the proposed quality estimation scheme by selecting an appropriate machine learner with good accuracy. After that, we conduct feature selection experiment by using random subset approach as proposed in [23], to evaluate the influence of the feature set we proposed.

Experimental setup We collect 1,000 items from Yahoo! Answers, Twitter, Foursquare, and Facebook: question answer pairs, tweets, tip messages, and posts. We label each item manually with a quality label in *good* and *bad* judgment. Two annotators conduct the annotation independently. When discrepancy happened, they would discuss to assign a final label. We then use 2/3 of the data for training and 1/3 for testing in the quality estimation experiments. The learned model is used to filter out the low quality content.

To obtain the relevance ground truth of the social media contents for the targeted cuisines, we pooled the top 20 results from various methods, namely, the vector space model, okapi BM25 model, language model and our proposed methods. We then asked the two annotators, who were not involved in the design of the proposed methods, to independently annotate whether the candidate social media item was relevant or not. When conflicts occurred, a third annotator was involved to make the final decision.

Table 3 Performance with one set of feature removed each time

System	All queries	inKB	NIL
Ours	0.886	0.879	0.914
w/o *Name-vs-Name*	0.792 (−10.5 %)	0.771 (−12.3 %)	0.878(−3.9 %)
w/o *Context-vs-Context*	0.829 (−6.3 %)	0.812 (−7.6 %)	0.899 (−1.6 %)
w/o *Geo-Context*	0.835 (−5.7 %)	0.823 (−6.4 %)	0.905 (−1.0 %)
w/o *Name-in-Context*	0.867 (−2.2 %)	0.857 (−2.5 %)	0.905 (−1.0 %)

The numbers in brackets are the difference ratio against the performance with the full set of features

Results and analysis To evaluate the most appropriate machine learning algorithm, we first conduct experiments using the proposed feature sets on a number of popular algorithms: support vector machines, logistic regression (LR), random forest (RF), sequential minimal optimization (SO), and voted perceptron (VP). Table 4 presents content quality estimation results of these algorithms. We can see that all machine learning algorithms have comparable accuracies in predicting the content quality, while SVM and logistic regression achieve the best accuracy and are significantly better than SO and VP.

To analyze the feature importance, we use the random feature selection method with correlation-based subset evaluation [23]. Figure 1 displays the most influential features. To compute the weights of each feature, we conduct a feature ranking evaluation by applying the Chi-squared statistic, with importance value normalized to between 0 and 100. For analysis purposes, we split our dataset into 75 % of train and 25 % test data randomly.

From Fig. 1, we can see that the most influential feature is the number of nonstop words in the text and the explicit and implicit votes. This result suggests that both the intrinsic features and the user interaction features are critical for estimating social media content quality. The leading importance of the nonstop word number indicates that informativeness is a good indicator. Moreover, these features are further strengthened by the readability scores.

5.2.2 *On Geo-enhanced content relevance*

Experimental setup To evaluate the effectiveness of the proposed Geo-enhanced content relevance model, we

Table 4 Accuracy of social media content quality estimation results with different classifiers

	SVM	LR	RF	SO	VP
Tweet	79.87	**83.32**	76.36	73.47	70.12
Y!A	73.43	**75.78**	70.63	68.92	66.63
Facebook	81.36	**83.27**	79.42	78.09	70.83
FourSquare	83.27	**84.74**	76.94	71.96	75.32

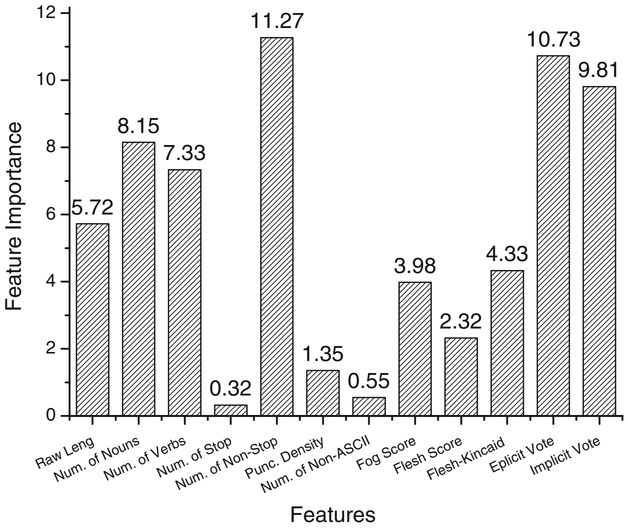

Fig. 1 The most influential general features for evaluating social medial content quality

randomly select 50 queries from the data in Table 1. Removing the duplicated cuisine queries, 31 queries on different cuisines are kept for this suite of experiments. We then form three search queries for each of the selected query, by expanding the original cuisine names with their name variants, the original context sentences, and the first sentence of the Wikipedia articles. The search queries are then submitted to Yahoo! Answers API, Twitter API, FourSquare API, and Facebook API. The returned documents are first filtered using the proposed quality evaluator. The high-quality items are then reranked using location information and diversified using MMR.

We compare with two baseline retrieval methods on reranking the contents under each social media source: (1) BM25, which ranks the documents by its BM25 score and (2) LM, which ranks the documents by the language modeling approach using Dirichlet-prior smoothing. We measure the performances using *nDCG@k*. Generally, *nDCG@k* normalizes $DCG@k = score_1 + \sum_{i=2}^{k} \frac{score_i}{log_2(i)}$, which is higher when more relevant documents are top ranked, where $score_i$ indicates the score of the *i*th ranked document.

Results and analysis: Overall, we can see from Table 5 that adding geocontext and diversification enhances the performances of the two baselines. For the two baselines, BM25 is better than LM: both models result in moderate performance on reranking the social media contents. With geocontext, both BM25 and LM achieve satisfactory results. With the divergence module providing the contextual information and geo-enhanced relevance, we can see significant improvements ($p < 0.05$) on all four levels of *nDCG*. These results suggest that for location-related relevance modeling, expanding the geocontext can be a plausible approach to combat the sparsity of meta information.

Now, we look at the breakdown of results on the four types of social media contents presented in Table 6. We can see that Yahoo! Answer QA pairs and Facebook messages achieve better results than tweets and Foursquare messages. One of the most distinct differences of the two groups are that the contents in the first group are longer than the second group. As the longer contents may contain more substantial information, this result may suggest that we need to consider more information to make the relevance model work better on social media content.

5.3 User study

As part of our evaluations, we implemented a local cuisine system test. Given a tourist's query and location, the system first finds the corresponding Wikipedia articles for the cuisine name in the query. It then searches and presents the aggregated social media contents reranked by the quality estimation module and the georelevance module. 12 international students who are not from Singapore are asked to give feedback on the usefulness of the information provided in the system. Each of them submit five to seven queries of Singapore cuisines mimicking the scenarios of their first visits to the city. Feedback on four metrics (1–10, the higher the better) are given by the students: the relevance of the knowledge base entries, the quality of the social media contents, the comprehensiveness of the social media contents, and the overall usefulness of the system. The average scores for the four metrics are 8.7, 7.9, 8.4, and 8.8, respectively. These results indicate that the system gives useful, informative, and relevant social media information about the local cuisines. Still, the quality needs some further improvements. One approach is to explore more social media sources to enlarge the scope that high quality contents can be sought. The other approach is to develop some auto editing function to rewrite the existing social media content available to the system.

Table 5 The *nDCG* performance of different methods in reranking relevant social media contents

Method	*nDCG@*			
	5	10	15	20
BM25	0.759	0.669	0.636	0.615
BM25 + Geo	0.787	0.701	0.663	0.648
BM25 + Geo + Div	**0.813**	**0.810**	**0.810**	**0.805**
LM	0.592	0.561	0.552	0.572
LM + Geo	0.723	0.684	0.636	0.615
LM + Geo + Div	**0.809**	**0.807**	**0.808**	**0.804**

Table 6 The *nDCG*@5 performance of the proposed methods in reranking relevant social media contents from different sources

Method	*nDCG@5*			
	Tweet	Y!A	Facebook	FourSquare
BM25 + Geo + Div	0.725	0.836	0.812	0.785
LM + Geo + Div	0.733	0.786	0.808	0.754

6 Conclusions and future work

Despite the importance and prevalence of social media as information sources for travelers, the research on increasing the accessibility of social media for tourist domain is sparse. In this work, we proposed to utilize the multi-source social media contents to resolve the local cuisines as one of the applications. Specifically, we proposed a general framework that consists of two major components. First, we built a location-aware linking module to resolve the pseudonym issue of the cuisine names. Second, we designed a location-aware aggregation module to compile the relevant social media contents in various media forms, given the resolved cuisine names. Experimental results showed the effectiveness of the two modules. The user study of the prototype system demonstrated that both the linking module and the aggregation module were helpful for tourists to understand the local cuisines. It remains an interesting future work to apply the proposed framework to various real-world problems where social media can help to enhance tourists' experiences in new localities.

References

1. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM, pp. 183–194 (2008)
2. Agirre, E., Chang, A., Jurafsky, D., et al.: Stanford-UBC at TAC-KBP. In: Proceedings of Test Analysis Conference 2009 (TAC09) (2009)
3. 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, ACM, New York, SIGIR '04, pp. 273–280 (2004). doi:10.1145/1008992.1009040
4. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In: Proceedings of ACL, Association for Computational Linguistics, Stroudsburg, pp 79–85 (1998). doi:10.3115/980845.980859
5. Bunescu, R., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: EACL, Trento, Italy, pp. 9–16 (2006)
6. Cao, Y., Xu, J., Liu, T.Y., Li, H., Huang, Y., Hon, H.W.: Adapting ranking SVM to document retrieval. In: Proceedings of SIGIR, ACM, Seattle, pp. 186–193 (2006)
7. Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H.: Learning to rank: from pairwise approach to listwise approach. In: Proceedings of the 24th International Conference on Machine Learning, ACM, Corvalis, Oregon, pp. 129–136 (2007)
8. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp. 335–336 (1998)
9. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, USA, CIKM '10, pp. 759–768 (2010). doi:10.1145/1871437.1871535
10. Cucerzan, S.: Large-scale named entity disambiguation based on wikipedia data. In: EMNLP-CoNLL, pp. 708–716 (2007)
11. Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, pp. 76–82 (2010)
12. Fleischman, M.B., Hovy, E.: Multi-document person name resolution. In: Proceedings of ACL-42, Reference Resolution Workshop (2004)
13. Gao, Y., Wang, M., Zha, Z.J., Shen, J., Li, X., Wu, X.: Visual-textual joint relevance learning for tag-based social image search. IEEE Trans. Image Process. **22**(1), 363–376 (2013)
14. Gao, Y., Wang, F., Luan, H., Chua, T.S.: Brand data gathering from live social media streams. In: Proceedings of International Conference on Multimedia Retrieval, ACM, p. 169 (2014)
15. Gottipati, S., Jiang, J.: Linking entities to a knowledge base with query expansion. In: Proceedings of EMNLP, Association for Computational Linguistics, Stroudsburg, pp. 804–813 (2011), http://dl.acm.org/citation.cfm?id=2145432.2145523
16. Han, X., Zhao, J.: Nlpr_kbp in tac 2009 kbp track: a two-stage method to entity linking. In: Proceedings of Test Analysis Conference 2009 (TAC09) (2009)
17. Han, X., Zhao, J.: Structural semantic relatedness: a knowledge-based method to named entity disambiguation. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 50–59 (2010)
18. Hoffart, J., Suchanek, F., Berberich, K., Weikum, G.: YAGO2: a spatially and temporally enhanced knowledge base from wikipedia. Artif. Intell. (2012)
19. Ji, H., Grishman, R., Dang, H., Griffitt, K., Ellis, J.: Overview of the tac 2010 knowledge base population track (2010)
20. Jiang, K., Liu, L., Xiao, R., Yu, N.: Mining local specialties for travelers by leveraging structured and unstructured data. Adv. Multimed. 15:15–15:15 (2012). doi:10.1155/2012/987124
21. Joachims, T.: Optimizing search engines using clickthrough data. In: KDD, pp. 133–142 (2002)

22. Kibble, R., Kibble, R., van Deemter, K., Deemter, K.V.: Coreference annotation: Whither? In: Proceedings of the 2nd International Conference on Language Resources and Evaluation, pp. 1281–1286 (2000)
23. Liu, H., Setiono, R.: A probabilistic approach to feature selection-a filter solution. In: ICML, Citeseer, vol. 96, pp. 319–327 (1996)
24. Mann, G.S.: Multi-document statistical fact extraction and fusion. PhD thesis, Baltimore, MD, USA, aAI3213760 (2006)
25. McNamee, P., Dang, H.T.: Overview of the TAC 2009 knowledge base population track. In: Proceedings of Test Analysis Conference 2009 (TAC09) (2009)
26. McNamee, P., Dredze, M., Gerber, A., Garera, N., Finin, T., Mayfield, J., Piatko, C., Rao, D., Yarowsky, D., Dreyer, M.: HLTCOE approaches to knowledge base population at tac 2009. In: Proceedings of Test Analysis Conference 2009 (TAC09) (2009)
27. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of CIKM, pp. 233–242. ACM, Lisbon (2007)
28. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of CIKM, ACM, Napa Valley, California, pp. 509–518 (2008)
29. Odijk, D., Meij, E., Graus, D., Kenter, T.: Multilingual semantic linking for video streams: Making "ideas worth sharing" more accessible. In: Proceedings of the 2nd International Workshop on Web of Linked Entities (WoLE 2013) (2013)
30. Si, L., Callan, J.: A statistical model for scientific readability. In: Proceedings of the 10th International Conference on Information and knowledge management, ACM, pp. 574–576 (2001)
31. Toba, H., Ming, Z.Y., Adriani, M., Chua, T.S.: Discovering high quality answers in community question answering archives using a hierarchy of classifiers. Inf. Sci. **261**, 101–115 (2014)
32. Varma, V., Bharat, V., Kovelamudi, S., Bysani, P., Santosh, G., Kumar, K., Maganti, N.: IIIT Hyderabad at TAC 2009. In: Proceedings of Test Analysis Conference 2009 (TAC09) (2009)
33. Xiang, Z., Gretzel, U.: Role of social media in online travel information search. Tour. Manag. 31(2),179–188 (2010). doi:10.1016/j.tourman.2009.02.016
34. Ye, M., Xiao, R., Lee, W.C., Xie, X.: On theme location discovery for travelogue services. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, SIGIR '11, pp. 465–474 (2011). doi:10.1145/2009916.2009980
35. Zhang, W., Su, J., Tan, C., Wang, W.: Entity linking leveraging automatically generated annotation. In: Proceedings of Coling, pp. 1290–1298. Beijing, China (2010)
36. Zheng, Z., Li, F., Huang, M., Zhu, X.: Learning to link entities with knowledge base. In: Proceedings of NAACL, Association for Computational Linguistics, Los Angeles, CA, pp. 483–491 (2010)
37. Zhu, X., Ming, Z.Y., Zhu, X., Chua, T.S.: Topic hierarchy construction for the organization of multi-source user generated contents. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, SIGIR '13, pp. 233–242 (2013). doi:10.1145/2484028.2484032