Using Latent Topic Features for Named Entity Extraction in Search Queries

Joe Polifroni and François Mairesse

Nokia Research Center Cambridge
4 Cambridge Center
Cambridge, MA 02142, USA
joseph.polifroni@nokia.com, francois.mairesse@nokia.com

Abstract

Search is one of the most quickly growing applications in the mobile market. As people rely more on portable devices for performing search, it becomes increasingly important to analyze user queries in order to achieve more targetted results over a broad set of search entities. While most previous work has relied on lexico-syntactic features and handcrafted knowledge sources, this paper investigates methods for learning latent semantic features from unlabelled user-generated content. We extract word-topic associations by training a Latent Dirichlet Allocation model on a corpus of online reviews, and show that this information improves named-entity classification performance over broad domain search queries. We believe that topical features provide a rich source of information from data with minimal manual effort, and no dependency on a specific language.
Index Terms: named entity extraction, spoken language processing, topic models, latent dirichlet allocation

1. Introduction

In 2009, the number of people with mobile broadband subscriptions worldwide surpassed those with fixed broadband subscriptions [1]. Using mobile devices for tasks normally associated with the internet, such as searches, is now commonplace. Voice is the obvious input modality for mobile search, given the size of mobile keyboards and the limitations of predictive text. Automatic Speech Recognition (ASR) has risen to the challenge by enabling applications where vocabulary size is essentially limitless and where users can say anything they want. The challenge now lies with natural language understanding to provide a deeper analysis of the words contained in the ASR transcription. In this paper, we describe our efforts to process search queries to determine both the generic *thing* being searched for (hereafter referred to as the *generic entity*) and the geographical location within which to constrain the search.

In processing search queries, we want to be able to identify search items at any level of specificity. For example, a user should be able to ask for "moo shu pork in San Francisco" or "PVC pipe in Cambridge" rather than "Chinese restaurants" or "hardware stores." Even when embedded in a spontaneous natural language query, the system should be able to extract both the location and the generic entity in question. Unfortunately, the set of possibilities for such generic entities is enormous, and structured lists and indices are not available for most business entities.

Hence we are interested in finding a data-driven and scalable solution that does not depend on handcrafted rules and could be ported to different languages. In this paper, we propose a way of using language found in reviews on consumer-oriented web sites as a way of identifying fine-grained search terms. Consumer-generated reviews can be seen as another type of spontaneously generated language. We hypothesize that the objects people refer to in reviewing business entities are the same ones that people search for when looking for businesses. To test this hypothesis, we train a Latent Dirichlet Allocation (LDA) model [2] over a large review corpus, from which we derive a set of features associating query words with underlying review topics. We use those features as part of maximum entropy classifier to make a three-way distinction among *generic entity*, *location*, and *unclassed word*. The output of this classifier is encoded as an FST with probabilities for each class. The Classifier FST is then composed with a statistical language model to yield an N-best list of named entity alignments. Figure 1 shows the system configuration. Our experiments compare topic features with a small set of features computed on the words themselves and on their left context [10]. Our results show that topic features improve performance, in support of the idea that unstructured data found on the web can be useful for spoken language understanding.

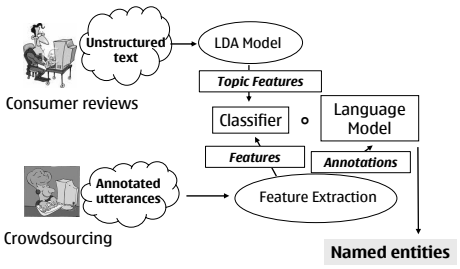

Figure 1: A diagram showing the configuration of the system, from crowdsourced/unstructured input to named entity output.

In the next section, we describe work related to the task we undertake here. In Section 3, we describe how we collected a corpus of search queries using crowdsourcing. Section 4 describes how the features based on LDA topics were derived in more detail. In Section 5, we describe the other features used by the classifier, as well as the language model used in the second stage of processing. Finally, Section 6 describes results on both topical feature evaluation and evaluation on the combined classifier and language model.

2. Related work

Although much research in named entity recognition has focused on written documents, the technology has also been applied to speech corpora, most notably broadcast news (e.g., [3]

inter alia). Bechet *et al.* [4] extract named entities from spontaneous speech within shorter utterances contained within the *HMIHY* corpus from AT&T. Huang *et al.* [5] and Jansche and Abney [6] perform named entity extraction on voicemail transcripts, by extracting "caller phrases," typically near the beginning of the message, in which the caller identifies him-/herself. Gotoh *et al.* [7] used a combination of a named entity tagger and a statistical language model to identify both words and categories from spoken input. In the work of Cucerzan and Yarowsky [8], a language independent approach to the problem was used, combining word-internal and contextual features. Recent work at AT&T, focusing on voice search, has also sought to extract locations from spoken input, along with query search terms [9].

In this work, we also look at extracting entities from textual utterances that are meant to mimic those users might speak to a mobile device. We make use of features that are associated with individual words and take context into account through a language model. In intent, our work most closely resembles that done at AT&T which focused on voice search [9]. Here, the intent was also to extract locations and query search terms from spoken input. This paper contrasts with previous work as well as our own work originally reported in [10] in two important ways. Firstly, the entities being recognized include unconstrained *generic entities* representing whatever object or service the user is looking for. This is arguably a more difficult task than the *location* and *proper name* entities classified in our original work, since we do not have ready resources such as gazetteers and census lists. Furthermore, features such as letter entropy scores do not distinguish the generic entities well, as they tend to be common words.

Secondly, identifying generic entities is helped by a novel set of features derived from a Latent Dirichlet Allocation (LDA) [2] model trained on unlabelled consumer-generated reviews. LDA is a popular technique for automatic discovery of semantic properties from consumer-generated review data [11]. It has also been applied as a first step in performing multi-aspect sentiment classification [13]. Here, we use LDA for shallow language understanding, to discover groupings of words within reviews and identify the entities of interest users refer to in search query utterances.

3. Creating the corpus

Due to the lack of availability of spoken language query data, we collected a corpus of search queries in written form generated by participants recruited through Amazon Mechanical Turk (AMT).

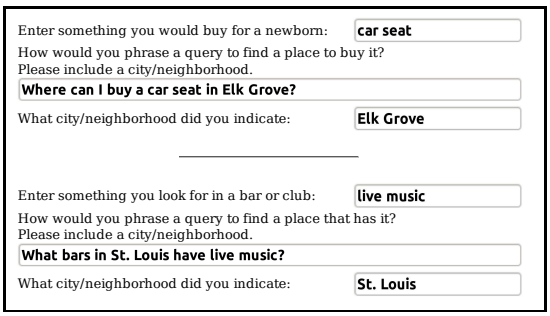

Figure 2: Data collection interface.

Figure 2 shows an example of a task each participant was asked to perform. The participants were first asked to identify a particular item of the sort that users might search for online. These items included a favorite dish at a particular type of restaurant, something someone might need on vacation, or an activity someone might do to relax on a weekend evening. Participants were then asked how they might phrase a spoken query to find such a thing, also indicating a city or neighborhood. They were also asked to indicate both the generic entity and geographical location they referred to in their query, in separate text boxes.

The resulting dataset consists of triples including a *generic entity*, a *location*, and an utterance containing both the entity and the location. We parsed these into data used for training and testing the algorithms described below, with each word marked automatically for the class we were interested in.

4. Topical feature extraction using LDA

Latent Dirichlet Allocation is a generative model widely used for unsupervised topic modelling, in which each document in the training data is associated with a multinomial distribution other K latent topics [2]. For each word w_n in document d, LDA assumes that a topic z_n is sampled from the topic distribution for d, and that w_i is sampled from the unigram word distribution for that topic. In order to limit overfitting and handle unseen words, the LDA model imposes a Dirichlet prior over the parameters of the topic and unigram distributions. The training process requires estimating both distributions given the observed documents, the fixed parameters of the Dirichlet priors, and the number of topics K. As an analytical solution is untractable, approximate inference methods such as Gibbs sampling or Variational Bayes are typically used. LDA can be seen as a discrete principal component analysis method for mapping a large feature vector (e.g., word counts of a document) into a more compact underlying representation (e.g., a distribution over topics).

This paper investigates the use of LDA for informing a broad-coverage named entity classification model, by making the following assumptions:

1. LDA can be used to extract topics of interest from large amount of consumer-generated media.
2. A word's topic distribution provides useful information for identifying named entities in unconstrained queries.

It is important to note that 'topic' here does not necessarily indicate a type of consumer product, e.g. knowing whether a word belongs to a 'topic' of positive sentiment adjectives could also help discriminate between entities.

The first step of our approach is therefore to extract topic probabilities for all unigrams in a large corpus of consumer-generated media. For each word to be classified in the user query, we then derive two types of features: (a) K topic distribution probabilities; and (b) the word's most likely topic index. The former feature set provides more fine-grained information than the latter, however it is not clear whether the low confidence scores produced by the LDA model are useful.

4.1. Experimental method

For our experiments, we train an LDA model on a corpus of 35,034 online user reviews of 5,153 venues/services in the Boston metropolitan area, spanning 73 distinct categories such as bakeries, sport shops, antiques, restaurants, car dealers, or

A	mexican food margaritas authentic margarita salsa restaurant cactus mex club mexico border tex chips tequila masa sangria mole tacos fajitas
B	dress dresses wedding bridal gown shop store alterations size fit gowns bride bridesmaids designer selection experience bridesmaid ana appointment
C	car auto repair shop work guys body fixed honest repairs cars parts mechanic damage insurance oil repaired tire rental engine
D	great recommend highly service experience excellent recommended friendly wonderful recently amp loved terrific location clean outstanding

Table 1: Most likely unigrams for 4 topics uncovered by LDA, out of 500 topics.

florists (2,941,455 words in total). Since we aim to answer unconstrained queries, the reviews cover a wide range of products and services, including shops, restaurants, hotels, doctors, mechanics, etc. We use the MALLET toolkit [14] to train an LDA model on this dataset and extract the K topic probabilities for each unigram in the reviews. To increase robustness to unseen words, unigrams that are not among the 500 most likely words for any topic are considered irrelevant and associated with a *null* topic. The training process uses Gibbs sampling with hyperparameter re-estimation every 10 iterations.

Table 1 illustrates the most likely words for four topics extracted using $K = 500$. One can see that despite not having access to labelled data, LDA learns a rich set of expression related to specific services, such as Mexican restaurants (topic A), wedding dresses (B) and garages (C). While query words belonging to product-specific topics are likely to be classified as named entities, LDA also learns topics characterizing other classes, e.g. positive affect words (topic D). We believe that both types of topics provide useful information for named entity classification.

5. Baseline named entity extraction system

We now summarize our named entity extraction system, including lexico-syntactic features as well as the language model used to rerank the output of our classifier.

5.1. Non-topical features

The classification process used a maximum entropy-based classifier. Output were three classes for each word, *name*, *location*, or *unclassed word*, with associated probabilities. In order to compare the effectiveness of the topical features, we build a baseline classifier using a subset of features calculated on either the word itself or on the word's left sibling. These features include:

- *stopWord*: presence in stopword list;
- *stopMinusOne*: presence of left-context word in stopword list;
- *fromBeginning*: distance from beginning of utterance;
- *fromEnd*: distance from end of utterance;
- *unigram*: word's unigram score calculated on the Google 1 T 5-gram corpus;
- *locP*: presence of word in location gazetteer;

We will refer to the set of features listed above as the *Original* set. An additional two features, comprised of the PoS for each word and for the word's left sibling, were derived using an off-the-shelf PoS Tagger [15]. In reporting results, the latter sets of features will be referred to as *PoS* and *Topics*.

where can i replace my car brakes in elk grove illinois
where can i *word* my *gen gen* in *loc loc loc*

Table 2: An example of an utterance from the training data and how this utterance was represented in the language model. The label *gen* is used for *generic entities* and *loc* for *locations*.

5.2. Language model

In order to exploit a larger utterance context, we use 80% of the search query corpus detailed in Section 3 to train a class-based language model in which each individual word is represented as either a *generic entity*, a *location* or an *unclassed word*. Any word that occured ten or more times in the training data is lexicalized. Table 2 shows an example of utterance from the training set and how they were represented in the language model.

As described in Section 1, the probabilistic output of the maximum entropy classifier was encoded as an FST, with arcs corresponding to each hypothesized class. The language model was also encoded as an FST, and the two FSTs were composed and the top choice chosen as the hypothesis for the utterance. Results on this processing are discussed in section 6.2.

6. Results

Before evaluating our full named entity extraction system, we tune our topical feature set on 80% of the search query corpus (8436 words) by focusing on the classifer only, leaving the remaining 20% for testing in Section 6.2.

6.1. Topical feature evaluation

The number of hidden topics K learned by the LDA model is an important parameter of our approach, as it determines both the granularity of individual topics and the number of topical features used for classification. While fine-grained topics provide more information—some of which might be discarded later on by the classifer—they also increase data sparsity issues. Hence we compare classification performance for different numbers of topics by doing a 10-fold cross validation on our training set. Figure 3 shows the word-level F-measure of the *generic entity* class. A first result is that the most likely topic feature on its own outperforms the topic distribution features on their own, suggesting that a hard topic assignment is beneficial. However, combining both types of topical features further improves performance, suggesting that the classifier benefits from knowing whether a word is associated with multiple topics. Interestingly, a paired T-test shows that with 500 topics LDA features significantly outperform PoS and *Original* features on their own (F=.75 vs F=.70 and F=.63, respectively, $p < .05$). When combining all features together, results show that adding topical features improves over our *Original* feature set, increasing the F-measure from .76 to .79 using 500 topics without PoS information (marginally significant at $p < .08$), and up to .80 with PoS (not significant). Interestingly, those results are obtained with only 3,767 words out of 8,436 being associated with a non-null topic. This suggests than more user-generated data would improve performance further.

Since the performance with 500 topics does not differ significantly from the performance with 200 topics despite the large feature space reduction, we use $K = 200$ for our final testing experiment in the next section.

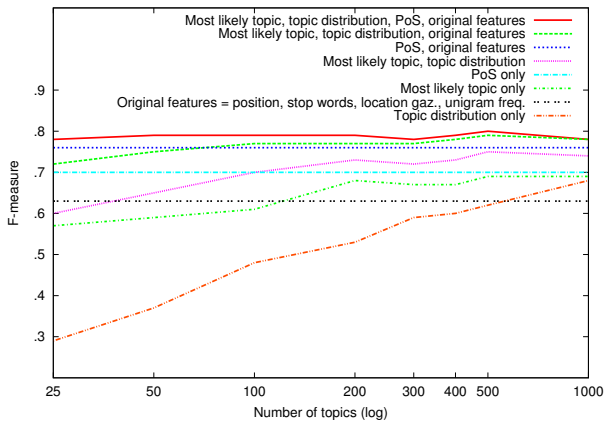

Figure 3: F-measure of the *generic entry* class on a 10-fold cross-validation for different numbers of topics and feature sets.

	Class.			Class.+LM		
Entities	Prec	Rec	F-meas	Prec	Rec	F-meas
Features: *Original*						
GenEnt	.70	.58	.63	.69	.72	.70
Loc	.75	.80	.78	.79	.88	.83
Features: *Original* + *Topics*						
GenEnt	.81	.73	.77	.79	.78	.78
Location	.88	.85	.87	.88	.94	.91
Features: *Original* + *Topics* + *PoS*						
GenEnt	.83	.76	.80	.80	.81	.81
Location	.88	.90	.89	.90	.94	.92

Table 3: The results on an 80/20 train/test set. On the left are the metrics applied to the output of the classifier alone; on the right shows results of the classifier in combination with a language model.

6.2. Evaluation on classifier + language model

For the evaluation of the classifier with the language model, we used the same training set as in the previous section (8436 words), and we use the remaining 20% of our corpus for testing (2100 words). Table 3 shows Precision, Recall, and F-measure for a subset of configurations of classifier features, with and without the language model. As can be seen, the best performance is achieved with a combination of the *Original* features, *PoS* features, and *Topics* features. However, it is interesting to note that performance with *Topics* improves dramatically over that with just the *Original* set of features, and is close to the performance with the combination of *Topics* and *PoS*. The system that used *Original* plus *Topics* was of particular interest, since that set of features is easily calculated from the words themselves or from unannotated, unstructured data. The language model requires annotation, but those annotations were available from crowdsourced tasks.

7. Conclusion and future work

In this paper, we have shown the effectiveness of features learned from LDA analysis of consumer-generated reviews for extracting named entities from user queries. There is an appeal to an approach that uses what people say about a particular entity, rather than structured data about business entities or keywords associated with webpages by developers. One hypothesis is that users speak about the same things in reviews that they ask about in web searches. Although this hypothesis remains to be tested, we find that user-consumer generated media provide useful information for broad-coverage named entity extraction. Furthermore, it is important to note that topical features can be trained for any language, without any handcrafting. In future work, we intend to make use of the topical features for response generation, by using the topic distribution associated with a query for producing a relevant summary of the most representative entities for a given topic.

8. References

[1] http://www.itu.int/ITU-D/ict/material/Telecom09_flyer.pdf

[2] Blei, D.M., Ng, A.Y., and Jordan, M.I., "Latent Dirichlet allocation", Journal of Machine Learning Research, 3, 993-1022, 2003.

[3] Horlock, J. and King, S., "Discriminative methods for improving named entity extraction on speech data", Proc., Eurospeech '03, Geneva.

[4] Béchet, F., Gorin, A., Wright, J., and Hakkani-Tür, D., "Named entity extraction from spontaneous speech in How May I Help You?", Proc, ICSLP '02, 2002.

[5] Huang, Jing, Zweig, Geoffrey and Padmanaghan, Mukund, "Information extraction from voicemail", Proc., Conference of the Association for Computational Linguistics (ACL), 2001.

[6] Jansche, M., and Abney, S., "Information extraction from voicemail transcripts", Proc. Conference on Empirical Methods in NLP, 2002.

[7] Gotoh, Y., Renals, S., and Williams, G., "Named Entity Tagged Language Models", Proc., ICASSP, 1999

[8] "Language Independent NER using a Unified Model of Internal and Contextual Evidence", Proc., CoNLL, 2002.

[9] Feng, J., Bangalore, S., and Gilbert, M., "Role of natural language understanding in voice local search", Proc. Interspeech, 2009.

[10] Polifroni, J., and Seneff, S., "Combining word-based features, statistical language models, and parsing for named entity recognition", Proc. Interspeech, 2010.

[11] Branavan, S. R. K. and Chen, H. and Eisenstein, J. and Barzilay, R., "Learning document-level semantic properties from free-text annotations", Journal of Artificial Intelligence Research, 34:11, 569-603, 2009.

[13] Titov, I., and McDonald, R., "Modeling online reviews with multi-grain topic models", Proc., WWW, 2008.

[14] McCallum, A. K., "MALLET: A Machine Learning for Language Toolkit", http://mallet.cs.umass.edu, 2002.

[15] Toutanova, K., Klein, D., Manning, C., and Singer, Y. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network", Proc., HLT-NAACL 2003.