Bootstrapping Domain Detection Using Query Click Logs for New Domains

Dilek Hakkani-Tür, Gokhan Tur, Larry Heck, Elizabeth Shriberg

Microsoft Speech Labs | Microsoft Research, Mountain View, CA

dilek@ieee.org, gokhan.tur@ieee.org, lheck@microsoft.com, elshribe@microsoft.com

Abstract

Domain detection in spoken dialog systems is usually treated as a multi-class, multi-label classification problem, and training of domain classifiers requires collection and manual annotation of example utterances. In order to extend a dialog system to new domains in a way that is seamless for users, domain detection should be able to handle utterances from the new domain as soon as it is introduced. In this work, we propose using web search query logs, which include queries entered by users and the links they subsequently click on, to bootstrap domain detection for new domains. While sampling user queries from the query click logs to train new domain classifiers, we introduce two types of measures based on the behavior of the users who entered a query and the form of the query. We show that both types of measures result in reductions in the error rate as compared to randomly sampling training queries. In controlled experiments over five domains, we achieve the best gain from the combination of the two types of sampling criteria.

Index Terms: Spoken language understanding, semi-supervised learning, web search queries, domain detection.

1. Introduction

Spoken language understanding (SLU) aims to obtain semantic analyses of spoken utterances [1]. Given an utterance, SLU in dialog systems extracts semantic information from the output of an automatic speech recognizer (ASR). The dialog manager (DM) then determines the next machine action given the SLU output. In the last decade, a variety of practical goal-oriented spoken dialog systems have been built for limited domains. Three key tasks in such targeted dialog and understanding applications are domain classification, intent determination and slot filling [2]. Domain classification is often completed first in SLU systems, serving as a top-level triage for subsequent processing. This modular design approach has the advantage of flexibility; specific modifications (e.g., insertions, deletions) to one domain class can be implemented without requiring changes to the other domains [3, 4]. Also, such an approach often yields more focused understanding in each domain, since the intent determination and slot filling only need to consider a relatively small set of classes over a single (or limited set) of domains.

Similar to intent determination, domain detection is often framed as a classification problem. More formally, given a user utterance or sentence x_i, the problem is to associate a set $y_i \subset C$ of semantic domain labels with x_i, where C is the finite set of domains covered. To perform this classification task, the class with the maximum conditional probability, $p(y_i|x_i)$ is selected:

$$\hat{y_i} = \underset{y_i}{\operatorname{argmax}}\, p(y_i|x_i)$$

Usually, supervised classification methods are used to estimate these conditional probabilities, and a set of labeled utterances is used in training. Collecting and annotating naturally spoken utterances to train these domain classifiers is often costly, representing a significant barrier to deployment, both in terms of effort and finances. However, it may be possible to overcome this hurdle by leveraging the abundance of *implicitly labeled* web search queries in search engines. Large-scale engines such as Bing or Google log more than 100 million search queries per day. Each query in the log has an associated set of URLs that were clicked on after the user entered the query. This user click information could be used to infer domain class labels and, therefore, to provide (noisy) supervision in training domain classifiers. For example, the queries of two users who click on the same URL (such as `http://www.hotels.com`) are probably from the same domain ("hotels," in this case). While users may sometimes click on the URLs randomly, it may be possible to detect on-target queries by investigating the click behavior of multiple users who search the same query. Furthermore, web search queries often represent keyword searches, such as "mountain view restaurant", which would be realized in natural conversations as complete utterances, such as "find me a restaurant near mountain view". For domain detection, lexical features (such as word n-grams of the input utterance) are typically the most informative classification features [5]. However, word n-grams extracted from natural language utterances and keyword search queries would not be the same: non-keywords are often missing in search queries, and keywords may be in a different order than in natural language utterances, requiring measures for sampling search queries that are in a form similar to that used for natural language utterances.

In this paper, we focus on bootstrapping domain detection models when a new domain is introduced to a spoken dialog application. The main contributions of this work are mining query click logs for in-domain examples to train classification models and selectively sampling queries using novel measures. We assume that the category of the clicked URL can be assigned as the domain label of the user query. For example, the label "hotel" is assigned to the user query "Holiday Inn and Suites" when the user has clicked on `http://www.hotels.com`. The biggest challenge in doing this is to find *natural language* queries with *high quality clicks*. To attack these problems, we propose a set of measures such as query frequency, target domain posterior probability, click entropy, domain-independent salient phrases, and syntactic parsing to sample queries and domain labels from the clicked URLs for use in domain detection.

In the next section, we first present related work and then describe the two types of measures we use for mining query click logs. Then, in Section 4, we present experiments on a set of natural language utterances from a spoken dialog system application using these methods.

28 – 31 August 2011, Florence, Italy

2. Related Work

Previous work on web search has benefited from the use of query click logs for improving query intent classification. For example, Li *et al.* [6] used query click logs to determine the intent of the web search queries (typically not in natural language) and infer the class membership of unlabeled queries from those of the labeled queries. They formed a bipartite graph of the queries and URLs the users clicked on, then transferred labels from queries to URLs and other queries using a label propagation algorithm [7, 8]. Note that these approaches focused on transferring labels without using the lexical content of the query. In our previous work, we proposed methods for integrating noisy supervision through click labels into the label propagation algorithm for semi-supervised learning of domain classifiers, where labels are propagated according to the lexical similarity of the queries and natural language utterances [9].

Recently, [10] studied several features, such as query entropy, dwell time and session length for mining high-quality clicks and showed that query entropy is the best single indicator for determining a successful click. In [11], Hassan et al. studied user action patterns and dwell time to estimate successful search sessions. We follow a similar approach, and use the posterior probability of the URL, query click entropy and frequency to sample high quality clicks in order to find queries to add as in-domain examples to the training set.

Regarding bootstrapping utterance classification models, two notable studies include the following: Di Fabbrizio *et al.* proposed reusing existing annotated data, especially for domain-independent intents and dialog acts, such as greetings [12]. Later, Chotimongkol *et al.* proposed using the predicate/argument structure of the utterances as extracted by Propbank-style shallow semantic parsing for bootstrapping intent determination models [13]. In the context of using data mined from the web to bootstrap dialog systems, the AT&T WebTalk system should be noted [14]. In that study, the goal is to retrieve the sentence from the website that best answers the user query. To the best of our knowledge, there is no previous study aiming to bootstrap utterance classification models using mined queries.

3. Approach

Query click data includes logs of search engine user queries and the links they click on from a list of sites returned by the search engine. Previous work has shown that click data can be used to improve search decisions [15]. However, most click data is very noisy and includes links that were clicked on almost randomly. We propose a set of measures, some of which have also been used in improving search, to sample queries and domain labels from the clicked URLs for use in domain detection. We form training data for the new domains from the sampled queries and add these to the labeled training data for other domains.

More specifically, we investigate two methods for sampling queries from these logs:

- noise filtering measures that use the distribution of clicks over the URLs aim to clean the noise coming in from erroneous clicks and sample only on-target queries. We try to estimate successful clicks by mining query click logs to gather the set of URLs clicked on by people who entered the exact same query.
- naturalness measures that aim to find natural language queries that would be uttered by users of a spoken dialog system. While query length is informative, we also use a syntactic parser to check if the query can be parsed as a grammatical sentence. To find the domain-independent salient phrases, we automatically construct dictionaries from naturally spoken utterances belonging to other domains.

3.1. Filtering Noise from Query Click Logs

We extract a set of queries from users who clicked on URLs related to our target domain categories, and then mine the query click logs to download all these search queries as well as the set of other links that were clicked on by search engine users who entered the same query. We use the following criteria to sample a subset of these queries to filter out the noisy queries:

- **Query Frequency:** refers to the number of times a query has been searched by different users in a given time frame. The motivation for using this feature is that in spoken dialog systems, users may ask the same things as web search users, hence adding frequent search queries to the domain detection training set may help to improve its accuracy.
- **Target URL posterior probability:** $P(U_t|q)$ refers to the probability that the users who type a query q click on the target URL, U_t, and is estimated as:

$$P(U_t|q) = \frac{F(U_t)}{\sum_{i=1}^{n} F(U_i)}$$

  where $U_i, i = 1, ..., n$ are the set of URLs clicked by the users of query q, and $F(U_i)$ is the number of times the URL U_i is clicked. The target URL is determined according to each domain, and can either simply be the base URL of the web site related to a domain (such as, `http://www.hotels.com`) or can also include alternative in-domain sites, for better coverage. The posterior probability aims to find on-target queries by sampling queries and assuming that if the $P(U_t|q)$ is high for a given query, then a click to the URL for this query is probably on-target and not a random click.
- **Query (Click) Entropy:** aims to measure the diversity of the URLs clicked on by the users of a query q, and is computed as

$$E(q) = -\sum_{i=1}^{n} P(U_i|q) ln P(U_i|q)$$

  Low click entropy may be a good indicator of the correctness of the domain category estimated from the query click label.

3.2. Measures Related to the Naturalness of the Query

Since most web search queries are one or two keywords, we employ the following naturalness criteria to sample a subset of these queries:

- **Query Length:** refers to the number of words in the query. The number of words in a query is usually a good indicator of NL utterances and search queries that include natural language utterances instead of simply a sequence of keywords that may be more useful as training data in SLU domain classification.
- **Syntax:** We use a syntactic parser (namely the Berkeley parser [16]) to check if the query can be parsed as a complete sentence. The main motivation for using syntax is to filter out keyword search queries.

Data Set	No. of examples	Avg. No. of words
Labeled training utterances	3,797	7.25
Labeled test utterances	1,014	7.14
Web search queries	2,171,021	4.21

Table 1: Data sets used in the experiments.

Experiment	Error Rate
No in-domain data	27.5%
All web queries	18.9%
Labeled in-domain data	6.2%

Table 2: Baseline experiments.

- **Domain-independent salient phrases:** Inspired by the How May I Help You (HMIHY) intent determination system [17], we find phrases which are salient for more than one domain. To this end, we use the available labeled training data from the other domains. For each n-gram n_j in this data set, we compute a probability distribution over domains: $P(domain_i|n_j)$, and then compute the KL divergence between this distribution and the prior probabilities over all domains $P(domain_i)$:

$$S(n_j) = KL(P(domain_i|n_j)||P(domain_i))$$

Then the word n-grams that show the least divergence from the prior distribution are selected as the domain independent salient phrases. These are phrases such as "show me all the" or "i wanna get information on" that frequently appear in natural language utterances directed to spoken dialog systems for information access. We check for the presence of such phrases in web search queries as an indicator of the naturalness of the query.

4. Experiments and Results

Similar to prior work on other utterance classification tasks, such as dialog act tagging [18] and intent determination [19], our approach relies on using icsiboost[1], an implementation of the AdaBoost.MH algorithm, a member of the boosting family of classifiers [20]. As features, we use word unigrams, bigrams and trigrams extracted from the training set. No feature normalization is performed to tag named entities (such as hotel or airline names in a travel system), as the system must learn the domain from the content words instead of from entity types since their annotation is typically non-trivial and noisy.

4.1. Data Sets and Experiment Set-up

We used a set of over 4,000 natural language utterances from a spoken dialog system application, as in our previous work [9]. These utterances belong to five different domain categories. Furthermore, we downloaded queries from Bing web search logs with the URLs clicked on by users. Table 1 shows the number of examples and the average utterance length in words in the data sets we used for the experiments.

For evaluating each method, we compute the error rate (ER), which is the number of examples for which the most probable domain category disagrees with manual annotation, divided by the total number of examples.

For controlled experiments, to see what happens when a new domain is introduced, we performed 5-fold experiments, where we remove one domain from the training set in each fold, and then add the mined in-domain data for the new domain. The average error rates from these 5 experiments are reported in the experiments in this section.

[1] http://code.google.com/p/icsiboost/

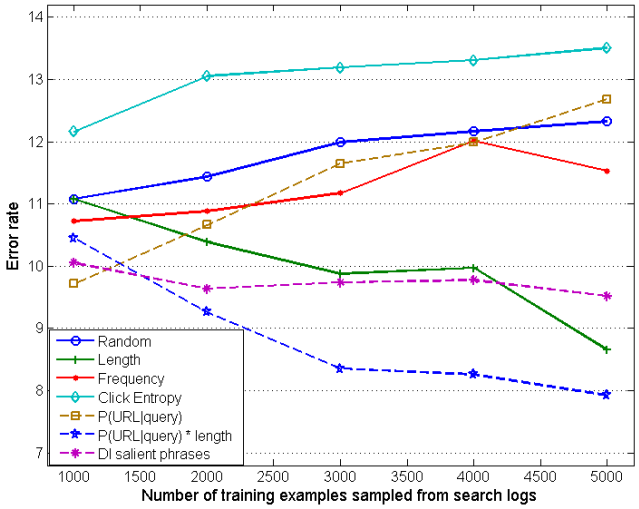

Figure 1: Learning curves of average error rates.

4.2. Baselines

Table 2 shows experiments with three baselines where: no labeled training data is used for the newly introduced domain, all web queries whose users click on URLs related to the new domain are added to the training set, and all the labeled natural language utterances collected from the spoken dialog application which are related to the newly introduced domain are added to the training set. Adding all queries mined from the search logs with class labels determined according to the clicked URL reduces the error rate from 27.5% with no in-domain queries to 18.9% on average over the five domains. The error rate of 6.2% is obtained when all manually labeled data is used.

4.3. Take-one-domain Out

Figure 1 shows the learning curves for average error rates (over 5 domains, due to space limitations) when 1000 to 5000 examples are selected using each measure and added to the training data set. In this plot, a random selection of queries added to the training can be viewed as a baseline. At all points, all measures except click entropy perform better than random sampling. When we examine the selected queries, click entropy includes several navigational queries where the search query is just the URL name, and such utterances are unlikely to appear in natural language queries in this type of applications. Query length and the presence of domain independent salient phrases in the query both significantly outperform random sampling of queries for training. When we use both posterior probability and query length in the sampling (by multiplying the two scores), we get queries that are on-target, resemble natural language utterances, and achieve the greatest reduction in error rate. When

Method	Sample size				
	1000	2000	3000	4000	5000
Random	11.08∓3.95	11.43∓2.59	11.98∓3.04	12.16∓3.01	12.32∓2.91
Length	11.08∓4.65	10.39∓3.41	9.88∓2.65	9.97∓3.00	8.65∓2.01
Frequency	10.72∓3.59	10.88∓3.68	11.18∓4.50	12.01∓4.34	11.53∓4.11
Click Entropy	12.17∓3.33	13.05∓4.22	13.19∓3.41	13.31∓3.59	13.51∓3.75
$P(URL\|query)$	9.72∓2.94	10.67∓4.65	11.65∓3.84	11.99∓3.61	12.68∓3.84
$P(URL\|query)*$ Length	10.45∓3.40	9.27∓2.61	8.36∓2.26	8.26∓1.86	**7.92**∓2.02
DI salient phrases	10.05∓3.88	9.64∓3.68	9.74∓3.80	9.78∓3.25	9.52∓3.07

Table 3: Learning curves of average error rates with standard deviations across domains.

5000 queries are added, the average error rate reduces from 12.3% for random sampling to 7.9% when the posterior probability and query length are used in selection. In this plot, we did not include results for when syntax is used, as this measure resulted in much higher error rates, mainly due to the mismatch between the sentences in the parser training data and web search queries.

When we look at the learning curves for each measure for different domains, we notice a different behavior. Table 3 shows the learning curves with standard deviation from the mean over the 5 domains. For example, even though query length by itself performs well on average, for some domains, such as search queries related to movies, books, and songs, this measure selects long descriptive queries, which result in higher error rates than random sampling. The higher standard deviation of the error rates when query length is used is an indicator of this behavior. Overall, posterior probability and length perform very well for all domains (except one, where it only slightly beats the performance of random sampling), and is fairly stable over examples.

5. Conclusions

We propose using web search query logs that include queries entered by users and the links they click on, to bootstrap domain detection for new domains. While sampling user queries from the query click logs to train new domain classifiers, we investigate two types of measures based on the behavior of the users who entered a query and the form of the query. We show that both types of measures result in reductions in the error rate, compared to randomly sampling the training queries. In controlled experiments over five domains, we achieve the best gain from the combination of posterior probability of the target URL given the query and query length, which mainly selects on-target natural language search queries.

Acknowledgments: We thank Ashley Fidler for her help with the manuscript.

6. References

[1] G. Tur and R. De Mori, Eds., *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, John Wiley and Sons, New York, NY, 2011.

[2] R. De Mori, F. Bechet, D. Hakkani-Tür, M. McTear, G. Riccardi, and G. Tur, "Spoken language understanding for conversational systems," *Signal Processing Magazine Special Issue on Spoken Language Technologies*, vol. 24, no. 3, pp. 50–58, May 2008.

[3] K. Komatani, N. Kanda, M. Nakano, K. Nakadai, H. Tsujino, T. Ogata, and H.G. Okuno, "Multi-domain spoken dialogue system with extensibility and robustness against speech recognition errors," in *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, Sydney, Australia, July 2006.

[4] C. Lee, S. Jung, S. Lee, and G.G. Lee, "Example-based dialog modeling for practical multi-domain dialog system," *Speech Communication*, vol. 51, pp. 466–484, 2009.

[5] G. Tur, D. Hakkani-Tür, and L. Heck, "What is left to be understood in ATIS?," in *Proceedings of the IEEE SLT Workshop*, Berkeley, CA, 2010.

[6] X. Li, Y.-Y. Wang, and A. Acero, "Learning query intent from regularized click graphs," in *Proceedings of SIGIR'08: the 31st Annual ACM SIGIR conference on Research and Development in Information Retrieval, Association for Computing Machinery, Inc.*, Singapore, July 2008.

[7] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Tech. Rep. CMU-CALD-02-107, CMU CALD technical report, 2002.

[8] X. Zhu, *Semi-Supervised Learning with Graphs*, PhD dissertation, Carnegie Melon University, 2005.

[9] D. Hakkani-Tür, L. Heck, and G. Tur, "Exploiting query click logs for utterance domain detection in spoken language understanding," in *Proceedings of the ICASSP*, Prague, Czech Republic, May 2011.

[10] A. Singla and Ryen W. White, "Sampling high-quality clicks from noisy click data," in *Proceedings of WWW2010*, Raleigh, North Carolina, USA, 2010.

[11] A. Hassan, R. Jones, and K. Klinkner, "Beyond DCG: User behavior as a predictor of a successful search," in *In the Proceedings of the ACM Conference on Web Search and Data Mining (WSDM 2010)*, New York City, USA, February 2010.

[12] G. Di Fabbrizio, G. Tur, and D. Hakkani-Tür, "Bootstrapping spoken dialog systems with data reuse," in *Proceedings of the SigDial Workshop*, Boston, MA, May 2004.

[13] G. Tur, D.Hakkani-Tür, and A. Chotimongkol, "Semi-supervised learning for spoken language understanding using semantic role labeling," in *Proceedings of the IEEE ASRU Workshop*, Puerto Rico, November 2005.

[14] J. Feng, S. Bangalore, and M. Rahim, "Webtalk: Mining websites for automatically building dialog systems," in *Proceedings of the IEEE ASRU Workshop*, U.S. Virgin Islands, December 2003.

[15] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior information," in *Proceedings of SIGIR*, Seattle, WA, USA, 2006, pp. 19–26.

[16] S. Petrov and D. Klein, "Learning and inference for hierarchically split PCFGs," in *Proceedings of the AAAI*, 2007.

[17] A. L. Gorin, G. Riccardi, and J. H. Wright, "How May I Help You?," *Speech Communication*, vol. 23, pp. 113–127, 1997.

[18] G. Tur, U. Guz, and D. Hakkani-Tür, "Model adaptation for dialog act tagging," in *Proceedings of the IEEE SLT Workshop*, 2006.

[19] P. Haffner, G. Tur, and J. Wright, "Optimizing SVMs for complex call classification," in *Proceedings of the ICASSP*, Hong Kong, April 2003.

[20] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.