TRANSLATING NATURAL LANGUAGE UTTERANCES TO SEARCH QUERIES FOR SLU DOMAIN DETECTION USING QUERY CLICK LOGS

Dilek Hakkani-Tür *Gokhan Tur* *Rukmini Iyer* *Larry Heck*

Speech Labs, Microsoft, Mountain View, CA

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

ABSTRACT

Logs of user queries from a search engine (such as Bing or Google) together with the links clicked provide valuable implicit feedback to improve statistical spoken language understanding (SLU) models. However, the form of natural language utterances occurring in spoken interactions with a computer differs stylistically from that of keyword search queries. In this paper, we propose a machine translation approach to learn a mapping from natural language utterances to search queries. We train statistical translation models, using task and domain independent semantically equivalent natural language and keyword search query pairs mined from the search query click logs. We then extend our previous work on enriching the existing classification feature sets for input utterance domain detection with features computed using the click distribution over a set of clicked URLs from search engine query click logs of user utterances with automatically translated queries. This approach results in significant improvements for domain detection, especially when detecting the domains of user utterances that are formulated as natural language queries and effectively complements to the earlier work using syntactic transformations.

Index Terms— domain detection, spoken language understanding, search query click logs, machine translation

1. INTRODUCTION

An important goal for spoken language understanding (SLU) in human/machine spoken dialog systems is to automatically identify the user's goal-driven intents for a given domain, as expressed in natural language, and extract associated arguments, or slots, according to a semantic template [1]. For multi-domain SLU systems, a top level domain detection, typically formulated as a classification problem, serves as a triage service. The state-of-the-art approach for training domain detection models relies on supervised machine learning methods that use lexical, contextual, and other semantic features. In this work, we introduce novel methods for translating naturally spoken user utterances into a form similar to keyword search queries, and extract features from search engine query click logs, related to the behavior of an abundance of users who typed in the same search query. Keyword search queries are typically shorter than natural language utterances and formed by phrase fragments rather than full sentences.

Enabling users to speak naturally to computers has been a goal for some time. Many spoken dialog systems motivate users to speak naturally by using explicit prompts, such as *You can speak naturally to me*. On the other hand, the success and broad use of keyword search engines imply the strength of keyword searches; some users attempt to speak in keywords, hoping for better machine understanding. Depending on whether they interact with a web search engine, another human, or an intelligent SLU system, users express their intents in different surface forms. Another motivation for this study is that, while it is difficult to formulate keyword searches for all user intents, a spoken dialog system should be able to handle both styles.

Search query click logs data includes past search engine users' queries and the links these users click from a list of sites returned by the search engine. Previous work has shown that click data can be used as implicit supervision to improve future search decisions [2, among others]. Regarding spoken language processing, in our previous work we mined data to train domain detection models when little [3] or no [4] in-domain data was available. Furthermore, we enriched the existing training data sets with new features, computed using the click distribution over a set of related URLs from search query click logs. Since the form of the natural language (NL) utterances differs from the shorter keyword search queries, to be able to match natural language utterances with search queries, we transformed the original utterances to query-like sentences using a syntax-based transformation and domain independent salient phrases learned from multi-domain user utterances [5]. While these transformations help improve domain detection, the transformed queries are not targeted to match the style of keyword search queries. In this work, we instead rely on statistical machine translation (SMT) between genres to convert user utterances to a search query form. Furthermore, for training the SMT models, we mine semantically similar natural language utterances and query pairs by walking on the bi-partite query click graph. Once the NL queries are translated into keyword queries, we extract features for domain detection, similar to [5].

Two novel contributions of this work is the mining of NL and keyword search query pairs from the search engine logs and use of statistical machine translation methods for the mapping, using this parallel corpus. The idea of translating utterances to the same language is not new, and is similar to paraphrasing by translation approach of Kok and Brockett[6] but is much stronger with the objective of translating from one genre of English to another.

The next section presents our approach to domain detection tsk, then Section 3 describes query click logs and mining of natural language utterances paired with keyword search queries, translation of NL utterances to a keyword search query, and extraction of features from search logs using the translated query. Section 4 presents the experiments and detailed results using a multi-domain SLU system. We conclude after a brief discussion of the results in Section 5.

2. DOMAIN DETECTION

In multi-domain SLU systems, domain classification is often completed first, serving as a top-level triage for subsequent processing. For example, the conversational system may support requests related to airline travel, weather, calendar scheduling, directory assistance, and so on.

Similar to intent determination systems like AT&T How May I Help You [7], domain detection is often framed as an utterance

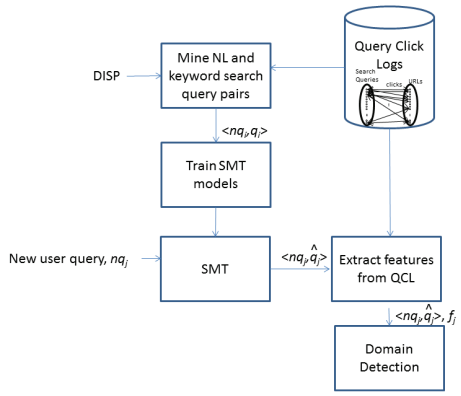

Fig. 1. A summary of the proposed approach to domain detection.

classification problem [3]. More formally, given a user utterance or sentence x_i, the problem is to associate a set $d_i \subset D$ of semantic domain labels with x_i, where D is the finite set of domains covered. To perform this classification task, the class with the maximum conditional probability, $p(d_i|x_i)$ is selected:

$$\hat{d}_i = \underset{d_i}{\operatorname{argmax}}\, p(d_i|x_i)$$

Usually, supervised classification methods are used to estimate these conditional probabilities, and a set of labeled utterances is used in training. Classification may employ lexical features such as word n-grams, contextual features such as the previous turn's domain, semantic features such as named entities in the utterance [8], syntactic features such as part-of-speech tags, topical features such as latent semantic variables [9] and so on.

3. APPROACH

The proposed approach relies on leveraging the implicitly annotated data coming from query click logs as additional features for training domain detection classification models. While this is straightforward in cases where a given user utterance is found in the query click logs with relatively high frequency, the language speakers use with an SLU system is different from typical queries. Note that, for some domains, such as one where the users are scheduling their own meetings, queries are unlikely to occur in the search logs, hence the absence of a query in the logs also provides information about the category of an utterance.

This study is motivated by the assumption that people typically have conceptual intents underlying their requests. They then generate different sequences of words depending on whether they interact with a web search engine, another human, or an intelligent SLU system. When they wonder about the *capacity of a Boeing 737*, they can form a simple query such as *capacity Boeing-737* when interacting with a search engine. The top wikipedia page will have the information requested. When they are interacting with an intelligent natural language dialog system, they can generate a more natural utterance, such as *what is the capacity of a Boeing 737 airplane*. In our previous work on sentence simplification [10], we proposed using a syntactic parsing based sentence transformation method to convert these input utterances to *capacity 737*, so that the classifier can perform better on them.

One immediate advantage we have noticed with this approach is that these transformed sentences look very much like search engine queries. Hence, it might be possible to use the URL click distributions for that query. For example, the utterance *I need to make a reservation for dinner* is transformed as *make reservation*, and that query may result in clicks to webpages like *opentable.com*. The domain classifier can exploit this orthogonal information in addition to the input utterance.

Our approach thus has four components, as depicted in Figure 2:

1. Learning a set of phrases that are typical of natural language utterances (we call them "domain independent salient phrases" (DISP), e.g., "I would like to"),
2. Mining semantically similar natural language query and keyword search query pairs using DISPs,
3. Training statistical machine translation models to convert NL utterances to keyword queries, and
4. Feature extraction from query click logs for domain detection.

The original and translated user utterances nq_j and q_j are checked against the query click logs, and a set of features, f_j are computed from the logs corresponding to them. If they are still not seen in the query click logs, this information is also provided to the classifier, as it indicates that the input probably belongs to a domain where there are no queries categorically related to information on the web.

In the following subsections we describe these key components.

3.1. Web Search Query Click Logs

Example clicks for some queries are shown below:

Query:	*who directed the count of monte cristo*
URL:	www.imdb.com/title/tt0047723/fullcredits
URL:	en.wikipedia.org/wiki/The_Count_of_Monte_Cristo
Query:	*zucca reviews*
URL:	www.yelp.com/biz/zucca-ristorante-mountain-view
URL:	reviews.opentable.com/0938/14689/reviews.htm

Note that each of the clicked links comes with frequencies showing the number of users entering that query clicked on that link. While in certain cases, the URL domain name is a direct indicator of the target domain (e.g., `opentable.com` receives queries about restaurant reservation, imdb.com receives queries about movies, etc.), general information web pages such as wikipedia.com provide only indirect information.

3.2. Domain-Independent Salient Phrases

Inspired by the How May I Help You (HMIHY) intent determination system [7], we find phrases that are salient for more than one domain. 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. To this end, we use the available labeled training data from other domains. For each n-gram n_j in this data set, we compute a probability distribution over domains $d_i \in D$: $P(d_i|n_j)$, and then compute the Kullback-Leibler (KL) divergence between this distribution and the prior probabilities over all domains $P(d_i)$:

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

Then the word n-grams that show the least divergence from the prior distribution are selected as the domain-independent salient phrases.

3.3. Mining Natural Language Utterance and Query Pairs

Search query click logs can be represented as a bi-partite graph, with two types of nodes, corresponding to queries and URLs. Left vertices correspond to queries and right vertices correspond to whole

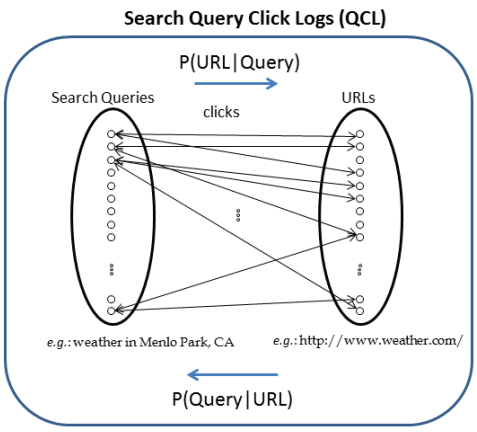

Fig. 2. The conceptual process for mining natural language user utterances paired with search queries from search query click logs.

URLs. An edge $e_{i,j}$ is added to the graph if a user who typed query q_i clicks on the URL u_j.

In this work, we search for search queries, nq_k, that include domain independent salient phrases. These represent natural language queries and form the seed set for mining pairs. Using the query click graph, we find a set of queries that are most semantically similar to the natural language queries. Similarity between an NL query, nq_k, and a query, q_i, is defined as:

$$sim(nq_k, q_i) = \sum_j P(q_i|u_j) \times P(u_j|nq_k)$$

This is similar to a two step walk on the query click graph.

However, since doing the walk for all possible URLs is very expensive, we first find the URL that has the maximum click probability given the utterance,nq_k:

$$\hat{u} = argmax_u P(u|nq_k)$$

Then, we approximate the similarity as:

$$sim(nq_k, q_i) = P(q_i|\hat{u}) \times P(\hat{u}|nq_k)$$

Then we use the pairs that have the highest similarity in training SMT models. Some of the pairs mined from Bing search engine logs are shown in Table 1. As seen, there are cases where the words or phrases in the input query are translated into other words (such as "biggest U S companies" is transformed into "fortune 500 companies"). We mined 30 million unique queries that include a DISP from the Bing search logs, and then walking through the click graph, we extracted 15 million NL and keyword query pairs.

3.4. Utterance-to-Query Translation

In order to train the natural language utterances to search queries, we employed the Moses statistical machine translation toolkit [11]. The training data consists of 1.7 million high precision pairs mined as described above. When a natural language query has more than one corresponding query based on the selected threshold, we tried using all examples or just the most frequent one, and found that the most frequent one results in a better match with keyword search queries.

Using Moses, we trained standard phrase-based machine translation models using the default settings. Note that this parallel data covers the whole web, and is not necessarily tuned to our target domains. The goal of this process is to provide a generic tool for translating input utterances. While it is clear that not all input utterances can be translated into keyword queries, even manually, the goal is to simulate the sort of translations provided in Table 1.

NL Query	**Keyword Query**
what are the signs of throat cancer	throat cancer symptoms
how many calories do i need in a day	calories per day
what are the biggest us companies	fortune 500 companies
are there any diet pills that actually work	diet pills that work
how do i know if i am anemic	anemic

Table 1. Sample natural language query and corresponding keyword search query pairs mined from QCL. DISPs extracted form the training set are italicized in the NL queries.

	Subset	No. Utt.	Avg. No. Words
Training	-	16,000	7.60
Development	-	2,000	7.65
Test	NL	1,243 (65.3%)	9.31
	Query	659 (34.7%)	4.27

Table 2. Data sets used in the experiments. NL refers to natural language subset, Query refers to Query-like utterances.

One key point worth noting is that the SMT model output by Moses needs to be tuned using in-domain data. To this end, we manually translated a held-out set of 2,000 utterances. About half of these utterances are judged to have some sort of query correspondences. This data is then fed into minimum error rate training (MERT), a tuning process well-known in the SMT community [12]. This step is shown to be essential to tune the weights of the system, as stylistically the pairs are very different than regular bilingual pairs.

Below we provide an example natural language utterance and the query we obtain using an SMT model and using syntactic transformation [5] to emphasize the difference between the previous work and current work:

> *NL utterance*: I want to make a reservation for dinner in sunnyvale
> *(a) Transformed utterance*: make reservation
> *(b) Translated utterance*: dinner reservation sunnyvale

3.5. Query Click Feature Extraction

Following the established literature on user utterance intent determination and domain detection, the baseline model uses lexical features, i.e., word n-grams as extracted from user utterances. In order to examine what users with similar intentions or information requests do with the web search results for their query, we search for each transformed utterance in our data set in the Bing web search query and click logs. We search all the queries in the training data set amongst the search queries, and download the list of clicked URLs and their frequencies. To reduce the number of features, we extract only the base URLs (such as opentable.com or wikipedia.com), as is commonly done in the web search literature. We use the list of the 1000 most frequently clicked base URLs for extracting classification features (QCL features). More formally, for each input user utterance, x_j, we compute $P(u_i|x_j)$, where u_i denotes the URLs and $i = 1, ..., 1000$.

4. EXPERIMENTS

4.1. Data Sets

In order to automatically detect the domain category of each utterance, we use both their word n-grams, and the base URLs clicked by

Approach	Overall ER	ER on NL Subset	ER on Query-like Subset	ER on subset with DISP	ER on subset without DISP
1: Word 1,2,3-grams (n-grams)	10.6%	11.3%	9.3%	9.9%	10.8%
2: n-grams + syntactic transformation (A)	9.4%	10.7%	6.8%	10.1%	9.1%
3: n-grams + SMT Approach (B)	9.3%	10.9%	6.2%	10.3%	8.9%
4: n-grams + (A) + (B)	8.5%	9.9%	5.8%	9.2%	8.2%

Table 3. Error rates when word n-grams as well as features computed from QCL are used for domain detection. syntactic transformation [5] indicates our earlier work using syntactic transformation and DISP removal.

search users who typed in the same query. We compile a dataset of user utterances from the users of a spoken dialog system. As mentioned earlier, some of these utterances are in the form of full conversational style natural language utterances (NL subset), for example, *I'd like to find out about weather in Mountain View tomorrow*, while others are more similar to web search queries, for example, *weather in Mountain View* (Query-like subset). We manually annotated the development and test set queries with style information. Table 2 shows the properties of the data sets and the (relative) frequencies of the two types of queries in each data set. While the average number of words per NL and query-like utterances is similar between the development and test sets, query-like utterances contain less than half the number of words as NL queries.

Each of the utterances in these data sets is manually labeled with one of 25 domain categories, that include both web related domains, such as *movies*, as well as others that usually do not appear in web search queries, such as *calendar*.

4.2. Results

Similar to prior work on other utterance classification tasks, such as dialog act tagging [13] and intent determination [14], our domain detection approach relies on using icsiboost[1], an implementation of the AdaBoost.MH algorithm, a member of the boosting family of discriminative classifiers [15].

To measure domain detection performance, we compute error rate (ER), that is the percentage of utterances that are not assigned the correct domain category, and F-measure, that is the harmonic mean of recall and precision.

Table 3 lists error rates when features computed from search query click logs are added to word n-grams as features. Similar to the previous set of results with syntactic transformations (ASRU-2011), using features from query click logs results in significant reductions in error rate (over 10% relative improvement over the baseline with both). Furthermore, when both the syntactic transformation and the output of the machine translation system are used, we observe significant improvements over both approaches (another 10% relative over the two approaches). The left four columns of Table 3 shows an analysis of error rates on two subsets of the test set that are manually and automatically (by checking presence of a DISP in the utterance) marked as NL and keyword search queries. The proposed approach improves error rates on both subsets.

5. CONCLUSIONS

We have presented a study using query click logs to better understand user utterances related to web, such as restaurant or movies. Two key novelties shown in this paper are mining natural language and keyword search query pairs and training a statistical machine translation system which can then be used to transform natural language input utterances. This approach resulted in significant reductions in the domain classification error rate and was shown to complement our earlier work based on syntactic transformation.

One important aspect of this study is that the implicit feedback extracted from query clicks provides an orthogonal view of the domain classification problem once user utterances are transformed into query language. This leads the way to many potential research ideas beyond this study, given the abundance of this contextual information. Note that one can also use search engine results after NL to keyword query translation for better understanding of web-related utterances.

Acknowledgments: We thank A. Celikyilmaz, D. Hillard, S. Parthasarathy and A. Fidler for helpful discussions.

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

6. REFERENCES

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

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

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

[4] D. Hakkani-Tür, G. Tur, L. Heck, and E. Shriberg, "Domain detection using query click logs for new domains," in *Proceedings of Interspeech*, Florence, Italy, 2011.

[5] D. Hakkani-Tür, G. Tur, L. Heck, A. Celikyilmaz, A. Fidler, D. Hillard, R. Iyer, and S. Parthasarathy, "Employing web search query click logs for multi-domain spoken language understanding," in *Proceedings of IEEE ASRU Workshop*, Waikoloa, HI, 2011.

[6] S. Kok and C. Brockett, "Hitting the right paraphases in good time," in *Proceedings of HLT-ACL*, Los Angeles, California, 2010.

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

[8] D. Hillard, A. Celikyilmaz, D. Hakkani-Tür, and G. Tur, "Learning weighted entity lists from web click logs for spoken language understanding," in *Proceedings of Interspeech*, Florence, Italy, 2011.

[9] A. Celikyilmaz, D. Hakkani-Tür, and G. Tur, "Multi-domain spoken language understanding with approximate inference," in *Proceedings of Interspeech*, Florence, Italy, 2011.

[10] G. Tur, D. Hakkani-Tür, L. Heck, and S. Parthasarathy, "Sentence simplification for spoken language understanding," in *Proceedings of the ICASSP*, Prague, Czech Republic, May 2011.

[11] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, and O. Bojar, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the ACL*, Prague, Czech Republic, 2007.

[12] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

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

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

[15] 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.