Detecting Out-Of-Domain Utterances Addressed to a Virtual Personal Assistant

Gokhan Tur, Anoop Deoras, Dilek Hakkani-Tür

Microsoft, USA

gokhan.tur@ieee.org, anoopd@microsoft.com, dilek@ieee.org

Abstract

Conversational understanding systems, especially virtual personal assistants (VPAs), perform "targeted" natural language understanding, assuming their users stay within the walled gardens of covered domains, and back-off to generic web search otherwise. However, users usually do not know the concept of domains and sometimes simply do not distinguish the system from simple voice search. Hence it becomes an important problem to identify these rejected out-of-domain utterances which are actually intended for the VPA. This paper presents a study tackling this new task, showing that *how* one utters a request is more important for this task than *what* is uttered, resembling addressee detection or dialog act tagging. To this end, syntactic and semantic parse "structure" features are extracted in addition to lexical features to train a binary SVM classifier using a large number of random web search queries and VPA utterances from multiple domains. We present controlled experiments leaving one domain out and check the precision of the model when combined with unseen queries. Our results indicate that such structured features result in higher precision especially when the test domain bears little resemblance to the existing domains.

Index Terms: conversational understanding, semantic parsing, keyword search, out-of-domain detection, machine learning, virtual personal assistants

10.21437/Interspeech.2014-69

1. Introduction

Spoken language understanding (SLU) in human/machine spoken dialog systems aims to automatically identify the domain and intent of the user as expressed in natural language (NL) and to extract associated arguments or slots [1] to achieve a goal. Most SLU tasks and approaches depend on the application and environment (such as mobile vs. TV) they have been designed for. Furthermore, in most multi-domain dialog systems [2, 3, 4, 5, among others], semantic processing is performed domain by domain (such as Calendar or Weather), instead of a global grammar or statistical model for all domains. Such "targeted" understanding also enables the system designers to decide on the capabilities of the envisioned system.

In such systems, the requested domain is determined either using an "acceptance" approach, i.e., each domain decides whether the utterance belongs to that domain, or using a "triage" approach, i.e., a top level classifier decides on the domain of the utterance, or both [6, 4, 7, among others]. The utterances which do not belong to any of the covered domains can simply be classified as "out-of-domain" using the classification confidence scores of virtual personal assistant (VPA) domain models [8]. Alternatively [9] proposed comparing the outputs of VPA domain models with a larger generic background model.

The problem with such a framework is that, this assumption only holds if the only purpose of the dialog system is to serve as a VPA, i.e., users only interact with the system when they have an unambiguous specific intent in their minds. Given the advances in voice search, especially with the boom of smart phones, the line between voice search and VPA has become very blurry. Now, it is not uncommon for VPA users to utter simple keywords (e.g., *onions hair* or *obama minimum wage*) hoping to get results from a back-off web search engine (such as in Microsoft Cortana or in Apple Siri backing off to Microsoft Bing).

Utterance	Domain
show me recent action movies by spielberg	movies
play me some romantic music	music
madonna like a virgin lyrics	music
obama politics	web search
yankees champion	web search
modern family hulu	web search
when is the yankees game tonight	orphan
watch modern family available on hulu	orphan
tivo the yankees game tonight	orphan
yankees score	orphan

Table 1: Example utterances in an entertainment VPA system. "Orphan" utterances have a specific unambigous intent which are not already covered

In any case, the conversational understanding field now has a previously unseen problem and that is of routing user utterance to one of the domains the user thinks is covered by the VPA or to web search. This problem is important not because we can have a top level classifier running as a preprocessing mechanism, but instead in order to understand whether there is a domain or intent which must have been covered (i.e., significant number of users utter requests in that domain to the VPA) but not covered either by design or due simply to ignorance. We call these out-of-domain utterances as "orphans", since in most cases, the generic web search will also not fulfill the users' requests. For example assume a VPA on entertainment which covers high traffic domains such as movies, music, and games. When the system is deployed and since the users do not necessarily know which domains are covered, they may have requests related to TV shows, such as when they are airing, whether the new episode is on, etc. Figure 1 gives example utterances in such a VPA system. The problem is then how to detect that these are out-of-domain utterances and may not be fulfilled by a search engine.

More formally, for us, the task is of building a classifier to detect *orphan* utterances using large amounts of utterances used to build domain specific models and random keyword queries hitting to web search. An orphan utterance is defined as whether it has a non-factoid *unambiguous specific intent* which is *known*

Utterance	*play me the trailer of avatar by james cameron*
Domain:	Movie
Intent:	Play_Trailer
Movie Name:	*Avatar*
Director:	*James Cameron*

Table 2: An example utterance with semantic annotations.

to be uncovered by any of the existing VPA domains. So, an utterance rejected by the domain models can either be an orphan or addressed to generic web search. It is irrelevant whether the web search engine can fulfill the user request as the capabilities of the search engines improve continuously. The classifier instead must return those utterances which could actually have been covered by a targeted language understanding system.

Significant ratio of utterances addressed to a VPA may be simple keywords. Similarly web search queries may include factoid questions in natural language form (e.g., *which is the tallest mountain in asia*). In that respect, this task is different than our earlier work which aims to find natural-language-like web search queries hitting to a target URL [10], but instead it is similar to the addressee detection or side-speech detection tasks [11] for conversational understanding. The utterances which are not already covered and are not orphan can then be processed by generic web search.

Once the orphan utterances are found, they can be processed online and/or offline, using a variety of methods, which are beyond the scope of this paper. The baseline would be the VPA system telling the user that his/her request is not covered yet, and/or a user experience scientist can analyze them to determine whether it is worth defining a new domain that covers these utterances. One can think of various semantic clustering methods to ease such a task.

In the next section we present the semantic parser that we employed in a representative VPA system. Then in Section 3 we present how we design features for building an orphan classifier. Section 4 presents experimental results.

2. Semantic Parsing

Typically, for targeted natural language understanding of machine-directed utterances, the task of semantic parsing is defined as extracting task specific arguments in a given frame-based semantic representation. Frame-based semantics is not a new concept, going back to 60s and 70s (e.g., DARPA Speech Understanding Research (SUR) project) [1]. For SLU systems, they are mostly motivated by the back-end capabilities of the system. Typically targeted semantic frames are designed to include domain and intent of the user and associated arguments (or slots). Figure 2 shows a semantic template for an example utterance in the music domain.

Since the intent and slots are very specific to the target domain and finding values of properties from automatically recognized spoken utterances may suffer from automatic speech recognition errors and poor modeling of natural language variability in expressing the same concept, spoken language understanding researchers employed known classification methods for filling frame slots of the application domain using the provided training data set and performed comparative experiments. These approaches used knowledge-based methods [2, 12, among others], probabilistic context free grammars [13]. However the state-of-the-art is using data-driven methods employing various machine learning methods [5].

3. Approach

Detecting uncovered utterances addressed to a VPA is surprisingly a hard task. It is more important to check how an intent is expressed more than understanding the specific intent. In that respect this task is more related to addressee detection [11] or dialog act tagging [14, 15] than domain detection task. For example if the utterance can be classified as a command (e.g., "send email to mom"), it is more likely to be addressed to the VPA than generic web search. Similarly if the content has only a named entity or a noun phrase and nothing else, it is more likely a keyword search (e.g., "hotel"), but of course not necessarily (e.g., "hotel reservation").

Based on these observations we have focused on 4 main types of features as described below. These features are extracted using the available VPA data and a random set of web search queries. Web search query set is noisy since some of the queries may have been meant to be addressed to a VPA (e.g., "weather in sunnyvale"). The lexical features show how far the use of generic word ngrams go. The syntactic and semantic features focus more on the structure of the input sentence than its content. The confidence scores from known domain classifiers are not used since the input of our system only consists of utterances which are *already rejected* by the covered domains.

The classifier of our choice is SVM [16]. We preferred a discriminative classifier, since they are less sensitive to the prior probability distribution compared to generative classifiers (like Naive Bayes). This is important in our case, since most of our training set consists of web search queries, while this is not the case during decoding. Furthermore, the feature space is very large considering all the word and POS tag ngrams, and SVMs are known to outperform other methods in binary classification especially for tasks with large sparse feature spaces [17, 18, among others].

The linear kernel SVM classification task can be more formally defined as follows: Given a collection of features extracted from VPA samples $VPA = \{(x_1, -1), ..., (x_m, -1)\}$ and web search query samples $Q = \{(x_{m+1}, 1), ..., (x_n, 1)\}$, forming the training data D, find the hyperplane, $\overline{w} \cdot \overline{x} - b = 0$, dividing these classes with the maximum margin.

3.1. Lexical Content

The lexical content is simply word n-grams in the input utterance. The intuition is that since the orphan classifier is trained with data from multiple domains, the classifier will not pick on the domain-specific content words (e.g., *cuisine* or *meal* for the restaurant domain), but instead on the domain-independent phrases (e.g., *could you please show me* or *what is the*). Since the distribution of words ngrams in web search queries will be very flat, such an approach provides a nontrivial baseline for the orphan classifier.

Table 3 provides the frequency of some VPA specific ngrams compared to web search dataset. When we check the most frequent non-stop words, VPA dataset has words related to covered domains, whereas the web search has words like *free*, *school*, *county*, and *sale*. The risk of using only lexical features is that, they may give higher confidences to queries which include some domain-dependent phrases, deteriorating the precision.

3.2. Syntactic Structure

Since we are not really interested in the content of the requests but instead on whether the utterance has a request such as *could*

Word	VPA Freq.	Web Search Freq.
me	0.69%	0.01%
i	0.45%	0.04%
my	0.34%	0.04%

Table 3: Relative frequencies of first person words in VPA and web search datasets.

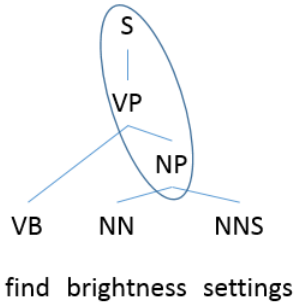

Figure 1: An example syntactic parse which can be converted into the structure feature of $S(VP(NP))$.

you please do this for me, as the next set of features, we propose using syntactic structure of the input utterances.

The baseline of structure features is nothing but part of speech (POS) tag ngrams. For example, if the first word's POS tag is a modal, MD, (e.g., "*could*") or a base form verb, VB, (e.g., "*play*"), it is a good indicator of a VPA utterance, compared to a proper noun, NNP. Similarly personal pronouns in base (PRP, e.g., "*I*") or genitive (PRP$, e.g., "*my*") forms may be good indicators.

Table 4 provides the frequency of most frequent POS tags of the first words for VPA and web datasets. As seen, VPA dataset has almost 10 times more words which are verbs to start a request, a strong indicator for detecting orphan utterances.

An extension of this pattern would be using the "shape" of the whole parse tree. Figure 1 provides an example of that with high level nodes indicating its shape. The syntactic parse tree of the sentence "*find brightness settings* can be converted into its nonterminals as $S(VP(NP))$, which is actually one of the most frequent shapes of VPA addressed utterances. One thing we have noticed is that, out of 100,000 example sentences, the web queries can be grouped into 1,346 shapes, while for VPA, this number is 20,829 due to the variance in natural language input, indicating that this feature is good mostly for recall than precision.

3.3. Semantic Structure

While syntactic parse features go a long way to capture the information beyond content, inspired from the original idea of targeted language understanding, semantic structure features are investigated. A typical semantic frame for the covered domains includes the "intent" of the users, which are typically in the shape of predicate/argument, such as *make/reservation, buy/ticket, play/trailer*, or *set/alarm* [19, 20, among others].

While it is not required for the utterance to explicitly have these predicates and intents (e.g., *what do people think of avatar* $\rightarrow$ *review/movie*), checking the existence of a predicate and a set of arguments in an utterance may be a strong feature for high precision orphan classification.

One can use a PropBank style shallow semantic parser [21] or a deeper one, such as a FrameNet parser [22]. While these

POS Tag	VPA Freq.	Web Search Freq.
VB	31.21%	3.01%
NNP	13.42%	54.27%
NN	5.72%	7.48%
WP	4.34%	1.57%
WRB	3.42%	2.47%
PRP	2.89%	0.37%
JJ	1.85%	8.66%

Table 4: Relative frequencies of top POS tags of utterance initial words in VPA and web search datasets.

parsers are not very robust with ASR output of naturally spoken utterances, most of the conversational understanding utterances are very short and simple to parse. In this study, we employed the Microsoft NLPWin parser [23], a generic knowledge-based semantic parser, which can output a semantic parse in the newly proposed AMR (abstract meaning representation) format [24].

For the example sentence above, one can get a semantic parse as below:

```
Input: find brightness settings
(f / find
      :ARG0 (y / you)
      :ARG1 (s / setting
            :mod (b / brightness))
      :mode imperative)
```

Mode shows the dialog act of the input utterance, like imperative, interrogative, or exclamation, if it is not a regular statement. $ARG0$ is usually the subject and $ARG1$ is the direct object, consistent with PropBank terminology [21]. "`mod`" is the modifier.

While one can add binary features such as whether there is a subject or object, similar to syntactic structure, we converted this structure into a semantic shape, dropping the lexical terminal nodes. For the example above, that would be $Pred(:\ ARG0\ :\ ARG1\ :\ mode\_imperative)$, which is actually the most frequent semantic shape for VPA addressed utterances. The most frequent pattern for web search queries is a stand alone concept (e.g., *facebook*) with a frequency of 30.6%. This figure is only 1.9% for VPA addressed utterances.

4. Experiments and Results

Instead of processing all available unseen data as orphan vs. web search for evaluation, we performed controlled experiments in this study, using the Microsoft Cortana setup. Using n-fold cross validation, we left one concrete domain out, as if it is not covered. Then the classifier is evaluated using how accurately the sentences of that domain can be picked when they are presented to the classifier along with a larger number of web search queries. This also guarantees that none of the known domains should claim these sentences.

The classification models are always trained using the covered domains except the left out one as the "VPA" class, and the web search queries as the "web" class. In this study we employed svmlight toolkit, using linear kernel with default parameters [1]. During training the 100K unique web search queries are picked from both head and mid frequency queries. Their frequencies are ignored as the head queries (e.g., *facebook* or *youtube*) would dominate the classifier otherwise. Table 6 summarizes the characteristics of the training and test data used in experiments.

[1] http://svmlight.joachims.org

Feature Used	Top sentences
Lexical	I need to get up an hour earlier tomorrow morning can you change the alarm this is going to happen every week set an alarm for tomorrow at 12:15 so I don't forget to get the kids in the car in time for the doctor appointment
POS Tags	I have to go to work 30 minutes early tomorrow, set my alarm to 30 minutes early create an alarm for weekdays to wake me up at 5 am
Syntactic Parse	wake me up at three o'clock what time do I need to wake up next week
Semantic Parse	set alarm for seven a.m. I no longer wish to hear the alarm

Table 5: Top selected sentences when the "alarm" domain is left out using different features.

	VPA	Web Search
Training	∼120K	100K
Test	∼20K	100K
Avg. # words	7.23	4.54

Table 6: Characteristics of the data used in experiments. VPA data belongs to 7 different domains, each with about 20K sentences on average. n-fold cross validation experiments are performed leaving one domain out at a time.

Table 5 provides examples from the highest scoring VPA examples for each of the feature types. It is immediately clear that models using syntactic and semantic parse structure features prefer shorter and crispier VPA examples, while models using lexical and POS tag features return longer sentences which have VPA specific key phrases such as *can you please* or *show me*. Since the model memorizes the content word ngrams from other domains, the lexical model also has key phrases such as *appointment* for Calendar domain or *don't forget* for Reminder domain.

The first set of results focus on "recall" only for the VPA class, i.e., checking the ratio of out-of-domain domain sentences classified correctly as VPA by the SVM classifier (without any thresholding). When there is no such orphan classifier in place, the uncovered utterances are handled by web search, giving a baseline recall value of 0. Table 7 presents results using 4 different types of features. The results are averaged for 7-fold experiments, one for each domain. The binary classification results are taken as is, with no thresholding on confidence. As seen, the lexical features are effective to distinguish the VPA sentences from web search queries. Even though there is little lexical overlap with the content words, the VPA indicator phrases such as "*can you*" or "*please*" are apparently good features for classification. The syntactic and semantic structure features on the other hand result in slightly better recall, marking more VPA sentences as VPA. Finally, using all features enabled the classifier to have perfect recall.

Of course, recall does not explain the whole picture. One needs to check precision as well using the test set of web search queries. However it is a non-trivial annotation task to mark 100K web search queries, and in most cases it is highly ambiguous. In order to solve this issue, as a second set of experiments, we instead checked precision at N (P@N) as the metric for evaluation: Top 100 web search queries which are confidently classified as VPA are manually checked for precision. The results presented in Table 8 show that while the lexical model has the highest hit rate, all of the selected non-factoid queries belong to already known domains. The POS tag based model relieves that problem with the expense of precision. The syntactic and semantic parsing based models outperform them with the highest ratios of out-of-domain non-factoid VPA addressed queries. The syntactic parsing based model suffered more from the factoid questions whose syntactic shape is exactly the same as VPA addressed queries (e.g., "*can you paint wood frame homes in winter*"). This is a nontrivial semantic disambiguation task. The distribution using all features resemble the one using only lexical features for factoids and in-domains, but results in the highest ratio of orphan queries.

Feature Set	Avg. Recall
Lexical	81.43%
POS Tags	85.14%
Syntactic Parse	85.57%
Semantic Parse	89.85%
All	100.00%

Table 7: Recall results: Ratio of correctly classified out-of-domain VPA utterances using various types of features.

Feature Set	Factoid	In-Domain	Out-of-Domain
Lexical	6%	50%	0%
POS Tags	17%	11%	4%
Syntactic Parse	48%	14%	11%
Semantic Parse	22%	4%	12%
All	4%	52%	18%

Table 8: Precision@100 results: Ratio of web search queries which are correctly classified as VPA, either from covered domains or uncovered domains.

5. Conclusions

We have presented a new classification task for conversational understanding: detecting whether the user has a request addressed to VPA or simple voice search. This task is important in order to capture and handle out of domain utterances in a VPA system for online of offline processing. Our results indicate that, *how* one utters a request is more important for this task than *what* is uttered, similar to addressee detection or dialog act tagging. In fact, using syntactic and semantic parse structure features resulted in better performance for recall and precision. Such a classifier can also be used to mine structurally similar queries or sentences from the web to bootstrap VPA models for new domains.

Future work will focus on efforts to automatically handle these out of domain utterances. These may include offline methods such as semantic clustering or online methods such as responding to user in an appropriate fashion (or both).

6. Acknowledgments

We would like to thank Geoff Zweig, Ruhi Sarikaya, and Larry Heck for many helpful discussions and Zhaleh Feizollahi for helping us with the datasets used in this study.

7. 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] J. R. Bellegarda, *Natural Interaction with Robots, Knowbots and Smartphones*. Springer, 2014, ch. Spoken Language Understanding for Natural Interaction: The Siri Experience.

[3] I. Lee, S. Kim, K. Kim, D. Lee, J. Choi, S. Ryu, and G. G. Lee, *Natural Interaction with Robots, Knowbots and Smartphones*. Springer, 2014, ch. A Two-Step Approach for Efficient Domain Selection in Multi-Domain Dialog Systems.

[4] 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 the IEEE ASRU*, Waikoloa, HI, 2011.

[5] G. Tur, Y.-Y. Wang, and D. Hakkani-Tür, *Computing Handbook, Third Edition*. Springer, 2014, ch. Understanding Spoken Language.

[6] M. Jeong and G. G. Lee, "Multi-domain spoken language understanding with transfer learning," *Speech Communication*, vol. 51, no. 5, pp. 412–424, 2009.

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

[8] I. Lane, T. Kawahara, T. Matsui, and S. Nakamura, "Out-of-domain utterance detection using classification confidences of multiple topics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 150–161, 2007.

[9] L. Heck and D. Hakkani-Tür, "Exploiting the semantic web for unsupervised spoken language understanding," in *In Prooceedings of the IEEE SLT Workshop*, Miami, FL, December 2012.

[10] A. Celikyilmaz, G. Tur, and D. Hakkani-Tür, "IsNL? A Discriminative Approach to Detect Natural Language Like Queries for Conversational Understanding," in *In Prooceedings of the Interspeech*, Lyon, France, August 2013.

[11] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and L. Heck, "Learning when to listen: Detecting system-addressed speech in human-human-computer dialog," in *In Prooceedings of the Interspeech*, Portland, OR, September 2012.

[12] W. Ward and S.Issar, "Recent improvements in the CMU spoken language understanding system," in *Proceedings of the ARPA HLT Workshop*, March 1994, pp. 213–216.

[13] S. Seneff, "TINA: A natural language system for spoken language applications," *Computational Linguistics*, vol. 18, no. 1, pp. 61–86, 1992.

[14] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[15] M. Core and J. Allen, "Coding dialogs with the DAMSL annotation scheme," in *Proceedings of the Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, Cambridge, MA, November 1997.

[16] V. N. Vapnik, *Statistical Learning Theory*. New York, NY: John Wiley and Sons, 1998.

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

[18] A. Moschitti, G. Riccardi, and C. Raymond, "Spoken language understanding with kernels for syntactic/semantic structures," in *Proceedings of the IEEE ASRU Workshop*, Koyoto, Japan, 2007.

[19] N. Gupta, G. Tur, D. Hakkani-Tür, S. Bangalore, G. Riccardi, and M. Rahim, "The AT&T spoken language understanding system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 213–222, 2006.

[20] G. Tur and L. Deng, *Intent Determination and Spoken Utterance Classification, Chapter 3 in Book: Spoken Language Understanding*. New York, NY: John Wiley and Sons, 2011.

[21] P. Kingsbury, M. Marcus, and M. Palmer, "Adding semantic annotation to the Penn TreeBank," in *Proceedings of the HLT*, San Diego, CA, March 2002.

[22] C. J. F. J. B. Lowe, C. F. Baker, "A frame-semantic approach to semantic annotation," in *Proceedings of the ACL - SIGLEX Workshop*, Washington, D.C., April 1997.

[23] G. E. Heidorn, *A handbook of natural language processing: Techniques and applications for the processing of language as text*. New York: Marcel Dekker, 2000, ch. Intelligent Writing Assistance, pp. 181–207.

[24] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider, "Abstract meaning representation for sembanking," in *Proceedings of the Linguistic Annotation Workshop*, 2014.