An Evaluation Framework for Natural Language Understanding in Spoken Dialogue Systems

Joshua B. Gordon[1] **and Rebecca J. Passonneau**[2]

[1]Department of Computer Science, Columbia University
[2]Center for Computational Learning Systems, Columbia University
{joshua, becky}@cs.columbia.edu

Abstract

We present an evaluation framework to enable developers of information seeking, transaction based spoken dialogue systems to compare the robustness of natural language understanding (NLU) approaches across varying levels of word error rate and contrasting domains. We develop statistical and semantic parsing based approaches to dialogue act identification and concept retrieval. Voice search is used in each approach to ultimately query the database. Included in the framework is a method for developers to bootstrap a representative pseudo-corpus, which is used to estimate NLU performance in a new domain. We illustrate the relative merits of these NLU techniques by contrasting our statistical NLU approach with a semantic parsing method over two contrasting applications, our CheckItOut library system and the deployed Let's Go Public! system, across four levels of word error rate. We find that with respect to both dialogue act identification and concept retrieval, our statistical NLU approach is more likely to robustly accommodate the freer form, less constrained utterances of CheckItOut at higher word error rates than is possible with semantic parsing.

1. Motivation

Natural Language Understanding (NLU) for Spoken Dialogue Systems (SDS) is complicated both by noise in automatic speech recognition (ASR), and by speech phenomena such as disfluencies, false starts, repairs and hesitations. ASR accuracy can degrade significantly in non-laboratory conditions. In (Raux, 2005), the authors observe a word error rate (WER) of 17% for native speakers in laboratory conditions which climbed to 60% in real world conditions. ASR error is compounded when the users comprise a diverse population of ages, accents, and varying prior experience using SDS. User utterances do not always conform to the vocabulary or grammar expected by the system. Background noise conditions further increase the difficulty of the recognizers' task.

Our SDS application is intended to handle library requests by telephone from an elderly population, thus we stand to benefit from an NLU framework that is particularly robust to ASR noise and caller disfluency. Our framework provides a mechanism to understand the benefits and limitations of two contrasting approaches to NLU for SDS, utterance classification and semantic parsing, which we considered as alternatives while designing our system. In future work, we will explore the utility of combining them in an ensemble approach. Our framework addresses the following questions:

- Is the NLU approach sufficiently robust to recognizer error and caller disfluency?
- How gracefully will NLU performance degrade with increasing WER?
- What is the maximum level of noise an NLU implementation can robustly accommodate?
- How does the impact of WER and choice of NLU vary across different application domains?

To facilitate exploring a new domain we provide a method to bootstrap a development corpus from a small, representative collection of user utterances. Our framework allows the experimenter to control for various parameters of the development corpus and NLU approach, including WER, and thus to estimate real-world performance under a variety of conditions. Our approach enables an offline exploration of the design tradeoffs in NLU faced by developers of information-seeking, transaction-based spoken dialogue systems when examining a new domain.

2. Related Work

To the best of our knowledge there is little related work with respect to offline estimation of NLU performance for a new domain. Our discussion of related work therefore addresses systems that rely on semantic parsing, classification, or mixed techniques to perform NLU.

There are numerous examples of dialogue systems designed to access a relational database. Applications include bus route information (Raux, 2005), restaurant guides (Johnston, 2002), weather (Zue, 2000) and directory services (Georgila, 2003). Many of the Olympus (Bohus, 2007) SDS upon which CheckItOut is based explicitly simplify the task of NLU by tightly constraining the set of allowable user utterances to include only the few words sufficient to retrieve the value of the desired attribute. Let's Go Public! (henceforth referred to as Let's Go), for instance, prompts users to say just the street name or neighborhood necessary for the database query. In aiming to elicit freer form utterances, CheckItOut is most similar in approach to (Gupta, 2006), which makes an explicit separation between the intent of the utterance, and the specific query terms contained within. The authors apply statistical utterance classification to broadly determine caller intent, followed by fixed rule based grammars to extract concepts.

Information seeking and transaction based dialogue systems typically perform natural language understanding on

ASR output before initiating a database query. Many techniques try to improve or expand ASR output. In CheckItOut and Let's Go, NLU is largely focused on determining which words of a possibly noisy utterance correspond to concepts in the domain database. For the directory service application in (Georgila, 2003) users spell the first three letters of surnames, and ASR results are expanded using frequently confused phones. In (Stoyanchev, 2009), a two-pass recognition architecture is applied to Let's Go to improve concept recognition in post-confirmation user utterances. To narrow the possible interpretations, decision trees have been used following a shallow semantic interpretation phase to classify utterances as relevant either to query type or to specific query slots (Komatani, 2005).

2.1. CheckItOut

CheckItOut is modeled on telephone library transactions at the Andrew Heiskell Braille and Talking Book Library, a branch of the New York Public Library and part of the National Library System. CheckItOut handles library book requests by telephone. It allows users the choice of requesting books by title, author or catalogue number. CheckItOut is primarily system-initiative, but allows users to choose whether to request a book by title, author or RC number. It seeks to elicit freer form, less constrained utterances on the part of the caller than is typical of a transactional dialogue system. We do not require the user to explicitly specify whether they are requesting a book by title or author, nor to limit their utterances to just the words contained in a book title. The bibliographic holdings of the Heiskell library include approximately 70,000 titles and 30,000 authors. The vocabulary for the 70 thousand book titles in our database is large, and has a high degree of overlap with other fields (e.g., author). The confusability of concepts is highlighted by figure 2, which describes the vocabulary overlap among unique words in the CheckItOut backend. Average book title length is 5.4 words (min=1; max=40); 26% of the titles are 1-2 words, 44% are 3 to 5 words, and 20% are 6 to 10 words. Differentiating between dialogue acts is challenging, and retrieving specific titles and authors from within user utterances is non-trivial.

2.2. Let's Go Public!

The Let's Go Public! bus information system is a telephone-based spoken dialogue system that provides access to bus route and schedule information in Pittsburgh. In a copy of a Let's Go corpus we were provided by Carnegie Mellon University, we calculate that the database contains about 70 bus routes, and 1300 place names in the greater Pittsburgh area. In order to provide bus schedule information, the system tries to identify the user's departure and arrival stop, and the departure or arrival time. Once the results are provided, the user can ask for the next or previous bus on that route, or can restart the conversation from the beginning to get information for a different route. The conversation begins with an open-ended "How can I help you?" prompt, but continues with a set of focused questions which the system asks in order to determine the departure place, arrival place and travel time. Let's Go is able to retrieve concepts contained in user responses in addition to those explicitly requested. For example, in response to the prompt requesting the departure place, users may respond also with the destination place and time. All concept values are confirmed either implicitly or explicitly by the system, depending on confidence.

2.3. Comparison and Design Tradeoffs

CheckItOut and Let's Go were selected for their significant disparity in mean utterance length (4.4 words for Let's Go vs. 9.1 for CheckItOut among unique utterances), in their vocabulary size (1825 for Let's Go vs. 6209 for CheckItOut), and database characteristics. In the development version of CheckItOut used for this experiment, our grammar was constructed from a subset of the 4000 most popular books in the full database. NLU in Let's Go employs Phoenix, a CFG semantic parser (Ward, 94) along with implicit and explicit confirm strategies to extract concepts from user utterances. In designing CheckItOut, we hypothesized that the NLU approach employed by Let's Go might prove inadequate to handle the freer form utterances of CheckItOut, particularly under higher levels of WER.

3. Comparative Evaluation Across Input Conditions

Throughout this paper we adopt the language of (Bangalore, 2006), which casts NLU in spoken dialog systems as a two stage process. The first task is determining the dialog act, or overall intent of an utterance. In the bus information domain, for example, a dialog act may correspond to a request for a departure time. The second task is to identify any domain specific concepts contained within the utterance. Concepts are named entities corresponding to a field in the database, such as a particular bus route, address, or departure time.

We contrast characteristics of semantic parsing and utterance classification as representative of two common NLU approaches for dialogue act and concept identification in spoken dialogue systems. Our framework operates by piping a corpus consisting of a series of utterances annotated for dialogue acts and concepts through alternative NLU frameworks. Annotations indicate the dialogue act of each utterance as well as substrings which correspond to specific concepts located in the database. For instance, the utterance "Do you have Melville's Moby Dick?", is annotated as a *Book Request* with a single author concept, *Melville*, and a single title concept, *Moby Dick*. We generate ASR of varying levels of WER via a simulation routine that borrows from both (Stuttle, 2004) and (Rieser, 2005), enabling us to evaluate NLU performance over changing levels of WER for different corpora, a key issue for dialogue system design.

3.1. Bootstrapping Development Resources

To illustrate our framework's capabilities, we compare a corpus collected by Let's Go to a bootstrapped development corpus modeled from transcripts of conversations between librarians and patrons that we recorded at the Andrew Heiskell Braille and Talking Book Library.[1] Here we

[1] A corpus of 82 transcribed calls will be released at the end of our project.

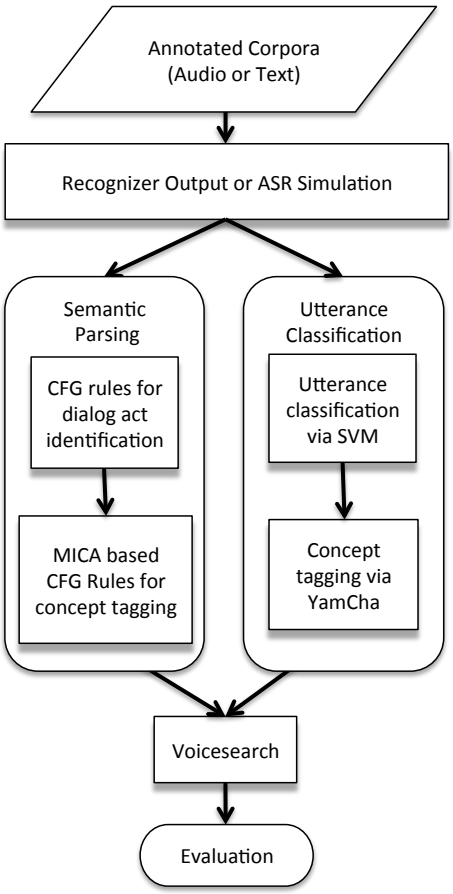

Figure 1: Architecture Overview

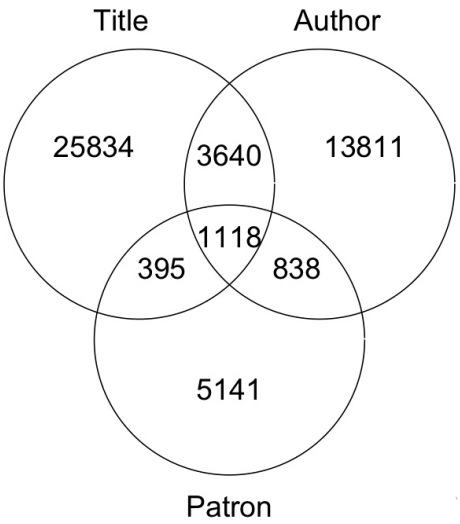

Figure 2: Vocabulary confusability among unique words in the database.

study a subset of utterances pertaining directly to book requests. Collections of these utterances are annotated for dialogue act and concept types, with two additional annotations highlighting what we refer to as the utterance *preamble* or *postamble*. For instance, the utterance "When does the next 61c depart?" is annotated as *Preamble:* "When does the next", *Bus concept*, *Postamble:* "depart?" To generate additional utterances, we substitute alternate preambles and postambles for the same concepts, and insert corresponding concepts from the database until a corpus of the desired size is reached. The generated utterances are representative of the language of the domain. We proceed by simulating speech recognition and disfluency over a desired range of WER.

3.2. Dialogue Act Identification

The challenge for syntactic and semantic analysis of spoken language is to achieve the right balance between accuracy and robustness to both the noise in speech recognition output, and the inherent discontinuities of spontaneous speech. It is not obvious how to optimally structure a robust grammar for spoken dialogue systems (Skantze, 2007). One alternative is to rely on robust semantic parsers in which the root nodes correspond to the dialogue act type, and where the terminals correspond to keyword sequences that directly match concepts. This is the approach taken in Let's Go, and in our baseline version of CheckItOut, both of which are implementations based on the Olympus/RavenClaw framework. Both systems use Phoenix (Ward, 94), a context free grammar (CFG) parser which outputs a semantic frame consisting of a possibly discontinuous sequence of slots when there are words that cannot be parsed. Each slot has an associated CFG; the root node is mapped to a concept. In our baseline CheckItOut, the RavenClaw-based dialogue manager takes concepts as input, and relies on a shallow plan hierarchy to infer the user's dialogue act type. The frame type plays only an implicit role in constraining the parse. Here, our aim is to compare two approaches to NLU for spoken dialogue systems, independent of the dialogue manager. Therefore, we treat the frame type assigned by Phoenix as the explicit representation of the dialogue act type.

Utterance classification casts dialogue act identification as a supervised learning problem. The major limitation of the statistical approach is the necessity of a training corpus which is often unavailable when exploring a new SDS application. If a corpus is available or can be constructed, however, the robustness to noise of the resulting system is typically high (Gupta, 2006). Good performance depends on design considerations, such as selecting the right granularity level for dialogue acts. Consider the utterance *"Do you have Melville's Moby Dick?"* for illustration. Too coarse a dialogue act label may be *book-request*, while an example of one too fine-grained may be *book-request-by-author-lastname*. Higher granularity increases misclassifications, while coarser granularity restricts the options presented to the dialogue manager. In constructing our classifiers, we cover a wide range of domain independent features, all of which utilize runtime information available to a production system and are computable in real time. Table 1 presents a subset of the lexical, syntactic, semantic, and acoustic features currently computed by the framework.

3.3. Concept Identification

Concepts refer to domain specific entities such as bus routes in Let's Go or titles in CheckItOut. Here we describe the se-

Classification Features	
AC	Acoustic Confidence
BW	Bag of Words (frequency weighted)
CW	Concept Word Distribution
CP	Cue Phrase N-grams (Webb, 2008)
LS	LSA Space Mapping
TF	TF / IDF Scores across Concepts
UP	Unigram POS Tag Distribution
UL	Utterance Length (phonemes)

Table 1: A subset of features computed for classification

mantic parsing approach to concept identification, followed by the classification approach. In both cases, the semantic interpretation phase includes voice search against the data base with the (possibly noisy) substrings presumed to correspond to concepts.

For semantic parsing of book titles, we begin with an initial hand-written Phoenix grammar to recognize distinct types of dialogue acts, such as book requests. The hand-written productions handle dialogue acts that contain no concepts or simple concepts such as phone numbers. They also specify the preambles and postambles of utterances of the form {preamble, concept, postamble}. To handle book title concepts, we automatically generate CFG productions from dependency parses of the full set of book titles we aim to cover. First we parse the book titles using MICA (Bangalore, 2009), a robust dependency parser, then with minor transformations, we map the dependency structures to CFG productions. This automatically builds linguistically motivated constraints on constituent structure and word order into the Phoenix productions. It also handles robustness to spoken language phenomena (e.g., false starts) and noisy recognition output through the feature of Phoenix parses that allows discontinuous parses. A book request frame for a book request by title consists of an optional preamble, a sequence of one or more title slots, and an optional postamble. Thus a book request parse of a noisy title string can consist of a sequence of title slots in which unparsable words are skipped. A post-processing phase concatenates the title slots within a single frame to represent a single book title, hence a single concept (see section 3.2. above).

I	WOULD	LIKE	THE	DIARY	A	ANY	FRANK	ON	TAPE
N	N	N	B_T	I_T	I_T	I_T	E_T	N	N

Figure 3: An ASR Hypothesis Tagged by YamCha.

For the classification approach, we cast concept identification as a named entity recognition problem in a manner similar to (Bangalore, 2006). The first phase involves segmenting the utterance so as to identify the substring corresponding to each concept. For this step, we employ YamCha (Kudo, 2003), a statistical tagger that learns using a support vector machine. YamCha is trained for each corpus and concept class (e.g., a book title or bus route), using a set of linguistic features over a sliding five word window. YamCha labels words within utterances as belonging to either the beginning, intermediate, or final word of a specific concept type, or as not belonging to a concept at all. Figure 3 shows an utterance labeled by YamCha. The identified concept is a title, *B_T*, *I_T*, and *E_T* correspond to the words predicted to begin, fall within, and end the title, respectively. In our experiment, the training data for CheckItOut is automatically labeled for supervised learning as our bootstrapped corpus is constructed. The training data for Let's Go was extracted from the annotated corpus provided by CMU.

3.4. Voice search

For both the semantic parsing and statistical machine learning approach, we use voice search (Wang, 2008) to query the database. Voice search refers to a partial matching database query operating on the phonetic level, where the search terms are the words identified by either the semantic parser or YamCha as belonging to a specific concept class. Voice search results are scored by Ratcliff / Obershelp similarity, which is the number of matching characters divided by the total number of characters in the string (Ratcliff, 1988). Matching characters are those in the longest common subsequence, then recursively in the longest subsequences in the unmatched regions. Figure 2 shows voice search results for a query on the title table of the Heiskell database. In scoring the NLU approaches studied in our framework we consider only the correctness of the highest ranked return. In previous work (Passonneau, 2010), we developed a machine learning approach to select the most likely candidate among these results or to ask question via a wizard of oz study.

Voice search results	
Anne Frank, the Diary of a Young Girl	.73
The Secret Diary of Anne Boleyn	.67
Anne Franke	.58

Table 2: Voice search results for search terms "the diary a any frank" extracted from the utterance in figure 3 by YamCha.

4. Results

When working with trigram based statistical language models, the short mean utterance length of Let's Go often results in a binary outcome: a dialogue act and its concepts are either recognized well, or not at all. In these instances, our results indicate NLU performance is highly correlated with WER regardless of the NLU technique. By comparison, the lengthier, less constrained utterances of CheckItOut provide a diverse feature set which enables recovery from higher WER with more complex NLU. Tables 3 and 4 describe the robustness of dialogue act and concept identification, respectively, for the studied corpora. The rapid decline in semantic parsing f-measure for dialog act identification is illustrative of the difficulty of writing a noise robust grammar by hand. By contrast, the first row of table 4 illustrates the comparative success of the MICA based dependency grammar. With respect to concept identification, YamCha is highly effective and robust to noise. We anticipate the WER of our deployed system to fall between .4

	$wer = .2$		$wer = .4$		$wer = .6$		$wer = .8$	
NLU	LetsGo	CIO	LetsGo	CIO	LetsGo	CIO	LetsGo	CIO
$Semantic\ Parsing$	.87	.58	.74	.36	.61	.30	.52	.23
$Statistical$	.73	.90	.69	.85	.65	.78	.55	.69

Table 3: Dialogue act identification (weighted f-measure) by classification technique

	$wer = .2$		$wer = .4$		$wer = .6$		$wer = .8$	
$concept$	CFG	Yamcha	CFG	Yamcha	CFG	Yamcha	CFG	Yamcha
$Title\ (MICA)$	.79	.91	.74	.84	.64	.70	.57	.59
$Author$	.57	.85	.49	.72	.40	.57	.34	.51
$Place$	.70	.70	.55	.53	.48	.46	.36	.34
Bus	.74	.84	.55	.65	.48	.46	.36	.44

Table 4: Voice search concept retrieval (f-measure) by search term extraction technique. Note that titles use a MICA dependency grammar, which yields substantially higher performance than manually generated rules.

and .6. Within this range, the combination of YamCha and Voice Search is effective in locating the correct title in 84% and 70% of instances, respectively.

5. Conclusion

We have presented an end-to-end evaluation framework which facilitates assessing the robustness to noise of NLU frameworks for information seeking, transaction based Spoken Dialogue Systems across varying WER and domains. We illustrate our framework's capabilities by comparing semantic parsing and utterance classification over two domains and four WERs. Our results indicate that richer NLU techniques can successfully accommodate the higher WER associated with less constrained user utterances in system-initiative systems.

Acknowledgements

This work was funded under the National Science Foundation under IIS-0745369.

6. References

Bangalore, Srinivas., et al. 2006. Introduction to the Special Issue on Spoken Language Understanding in Conversational Systems. Speech Communication 48(3-4): 233-238.

Bangalore, Srinivas; Boulllier, Pierre, B., et al. 2009. MICA: a probabilistic dependency parser based on tree insertion grammars application note. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Boulder, Colorado.

Bohus, D; Raux, A; Harris, T; Eskenazi, M; Rudnicky, A. 2007. Olympus: an open-sourceframework for conversational spoken language interface research. in Proceedings of HLT-2007. Rochester, NY.

Georgila, Kallirroi; Sgarbas, Kyrakos; Tsopanoglou, Anastasios; Fakotakis, Nikos; Kokkinakis, George. 2003. A speech-based human-computer interaction system for automating directory assistance services.International Journal of Speech Technology, Special Issue on Speech and Human-Computer Interaction, 6(2), 145-159.

Gupta, N., G; Tur, et al. 2006. The AT&T spoken language understanding system. Audio, Speech, and Language Processing, IEEE Transactions on 14(1): 213-222.

Hahn, S; Lehnen, P; Raymond, C; Ney, H. 2008. A comparison of various methods for concept tagging for spoken language understanding, in Proc. of the Sixth Int. Conf. on Language Resources and Evaluation (LREC), Marrakech, Morocco.

Johnston, Michael; Bangalore, Srinivas; Vasireddy, Gunaranjan; Stent, Amanda; Ehlen, Patrick; Walker, Marilyn A., et al. 2002. MATCH - An architecture for multimodal dialogue systems. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 376-383.

, Komatani, Kazunori; Kanda, Naoyuki; Ogata, Tetsuya; Okuno, Hiroshi G. 2005. Contextual constraints based on dialogue models in database search task for spoken dialogue systems. The Ninth European Conference on Speech Communication and Technology (Eurospeech), pp. 877-880.

Passonneau, Rebecca J.; Epstein, Susan; Ligorio, Tiziana; Gordon, Joshua; Bhutada, Pravin. 2010. Learning about voice search for spoken dialogue systems. To appear, 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010). Los Angeles. June 1-6, 2010.

Ratcliff, J. W. and D. Metzener 1988. Pattern Matching: The Gestalt Approach, Dr. Dobb's Journal.

Raux, A; Langner, B; Black, A; Eskenazi, M. 2005. Let's Go Public! Taking a spoken dialog system to the real world. Paper presented at the Interspeech 2005 (Eurospeech), Lisbon, Portugal.

Edlund, J; Skantze, G; Carlson, R. 2004. Higgins: a spoken dialogue system for investigating error handling techniques, In Proceedings of ICSLP.

Skantze, G. 2007. Error Handling in Spoken Dialogue Systems: Managing Uncertainty, Grounding and Miscommunication. Doctoral Thesis, KTH, Stockholm, Sweden.

Stoyanchev, Svetlana; Stent, Amanda. 2009. Predicting concept types in user corrections in dialog. Proceedings of the EACL Workshop SRSL 2009, the Second Work-

shop on Semantic Representation of Spoken Language, pp. 42-49.

Stuttle, Matthew; Young, Steve. 2004. A Framework for Dialogue Data Collection with a Simulated ASR Channel. In Proceedings of the ICSLP, Jeju, South Korea.

Kudo, Taku; Matsumoto, Yuji. 2003. Fast methods for kernel-based text analysis. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1. Sapporo, Japan.

Webb, Nick; Liu, Ting. 2008. Investigating the portability of corpus-derived cue phrases for dialogue act classification. Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1. Manchester, United Kingdom.

Wang, Ye-Yi; Yu, Dong; Ju, Yun-Cheng; Acero, Alex. 2008. An Introduction to Voice Search, IEEE Signal Processing Magazine: Special Issue on Spoken Language Technology.

Zue, Victor; Seneff, Stephanie; Glass, James; Polifroni, Joseph; Pao, Christine; Hazen, Timothy J., et al. 2000. A Telephone-based conversational interface for weather information. IEEE Transactions on Speech and Audio Processing, 8, 85-96.

Ward, Wayne and Sunial Issar. 1994. Recent improvements in the CMU spoken language understanding system. Paper presented at the ARPA Human Language Technology Workshop, Plainsboro, NJ.

Rieser, Verena; Kruijff-Korbayov'a, Ivana; Lemon, Oliver. 2005. A corpus collection and annotation framework for learning multimodal clarification strategies. Paper presented at the Proceedings of the Sixth SIGdial Workshop on Discourse and Dialogue.