Using Out-of-Domain Data for Lexical Addressee Detection in Human-Human-Computer Dialog

Heeyoung Lee[1*] **Andreas Stolcke**[2] **Elizabeth Shriberg**[2]
[1]Dept. of Electrical Engineering, Stanford University, Stanford, California, USA
[2]Microsoft Research, Mountain View, California, USA
heeyoung@stanford.edu, {anstolck,elshribe}@microsoft.com

Abstract

Addressee detection (AD) is an important problem for dialog systems in human-human-computer scenarios (contexts involving multiple people and a system) because system-directed speech must be distinguished from human-directed speech. Recent work on AD (Shriberg et al., 2012) showed good results using prosodic and lexical features trained on in-domain data. In-domain data, however, is expensive to collect for each new domain. In this study we focus on lexical models and investigate how well out-of-domain data (either outside the domain, or from single-user scenarios) can fill in for matched in-domain data. We find that human-addressed speech can be modeled using out-of-domain conversational speech transcripts, and that human-computer utterances can be modeled using single-user data: the resulting AD system outperforms a system trained only on matched in-domain data. Further gains (up to a 4% reduction in equal error rate) are obtained when in-domain and out-of-domain models are interpolated. Finally, we examine which parts of an utterance are most useful. We find that the first 1.5 seconds of an utterance contain most of the lexical information for AD, and analyze which lexical items convey this. Overall, we conclude that the H-H-C scenario can be approximated by combining data from H-C and H-H scenarios only.

1 Introduction

Before a spoken dialog system can recognize and interpret a user's speech, it should ideally determine if speech was even meant to be interpreted by the system. We refer to this task as addressee detection (AD). AD is often overlooked, especially in traditional single-user scenarios, because with the exception of self-talk, side-talk or background speech, the majority of speech is usually system-directed. As dialog systems expand to more natural contexts and multiperson environments, however, AD can become a crucial part of the system's operational requirements. This is particularly true for systems in which explicit system addressing (e.g., push-to-talk or required keyword addressing) is undesirable.

Past research on addressee detection has focused on human-human (H-H) settings, such as meetings, sometimes with multimodal cues (op den Akker and Traum, 2009). Early systems relied primarily on rejection of H-H utterances either because they could not be interpreted (Paek et al., 2000), or because they yielded low speech recognition confidence (Dowding et al., 2006). Some systems combine gaze with lexical and syntactic cues to detect H-H speech (Katzenmaier et al., 2004). Others use relatively simple prosodic features based on pitch and energy in addition to those derived from automatic speech recognition (ASR) (Reich et al., 2011).

With some exceptions (Bohus and Horvitz, 2011; Shriberg et al., 2012), relatively little work has looked at the human-human-computer (H-H-C) scenario, i.e. at contexts involving two or more people who interact both with a system and with each other.

[*] Work done while first author was an intern with Microsoft.

Proceedings of NAACL-HLT 2013, pages 221–229,
Atlanta, Georgia, 9–14 June 2013. ©2013 Association for Computational Linguistics

Shriberg et al. (2012) found that novel prosodic features were more accurate than lexical or semantic features based on speech recognition for the addressee task. The corpus, also used herein, is comprised of H-H-C dialog in which roughly half of the computer-addressed speech consisted of a small set of fixed commands. While the word-based features map directly to the commands, they had trouble distinguishing all other (noncommand) computer-directed speech from human-directed speech. This is because addressee detection in the H-H-C scenario becomes even more challenging when the system is designed for natural speech, i.e., utterances that are conversational in form and not limited to command phrases with restricted syntax. Furthermore, H-H utterances can be about the domain of the system (e.g., discussing the dialog task), making AD based on language content more difficult. The prosodic features were good at both types of distinctions—even improving performance significantly when combined with true-word (cheating) lexical features that have 100% accuracy on the commands. Nevertheless, the prior work showed that lexical n-grams are useful for addressee detection in the H-H-C scenario.

A problem with lexical features is that they are highly task- and domain-dependent. As with other language modeling tasks, one usually has to collect matched training data in significant quantities. Data collection is made more cumbersome and expensive by the multi-user aspect of the scenario. Thus, for practical reasons alone, it would be much better if the language models for AD could be trained on out-of-domain data, and if whatever in-domain data is needed could be limited to single-user interaction. We show in this paper that precisely this training scenario is feasible and achieves results that are comparable or better than using completely matched H-H-C training data.

In addition to studying the role of out-of-domain data for lexical AD models, we also examine which words are useful, and how soon in elapsed time they are available. Whereas most prior work in AD has looked at processing of entire utterances, we consider an online processing version where AD decisions are to be made as soon as possible after an utterance was initiated. We find that most of the addressee-relevant lexical information can be found in the first 1.5 seconds, and analyze which words convey this information.

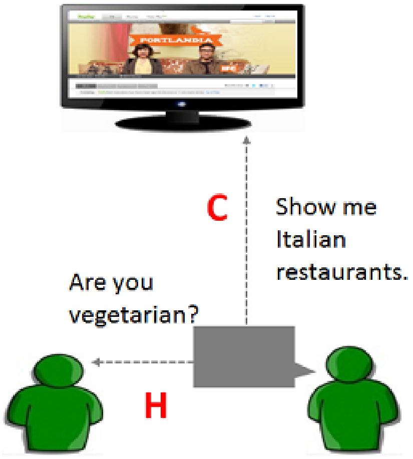

Figure 1: Conversational Browser dialog system environment with multi-human scenario

2 Data

We use in-domain and out-of-domain data from various sources. The corpora used in this work differ in size, domain, and scenario.

2.1 In-domain data

In-domain data is collected from interactions between two users and a “Conversational Browser” (CB) spoken dialog system. We used the same methodology as Shriberg et al. (2012), but using additional data. As depicted in Figure 1, the system shows a browser on a large TV screen and users are asked to use natural language for a variety of information-seeking tasks. For more details about the dialog system and language understanding approach, see Hakkani-Tür et al. (2011a; 2011b).

We split the in-domain data into training, development, and test sets, preserving sessions. Each session is about 5 to 40 minutes long. Even though the whole conversation is recorded, only the segments captured by the speech recognition system are used in our experiments. Each utterance segment belongs to one of four types: computer-command (C-command), comprising navigational commands to the system, computer-noncommand (C-noncommand), which are computer-directed utterances other than commands, human-directed (H), and mixed (M) utterances, which contain a combina-

Table 1: In-domain corpus

(a) Sizes, distribution, and ASR word error rates of in-domain utterance types

Data set	Train	Dev	Test	WER
Transcribed words	6,490	11,298	9,486	
ASR words	4,649	6,360	5,514	59.3%
H (%)	19.1	48.6	37.0	87.6%
C-noncomm. (%)	38.3	27.8	32.2	32.6%
C-command (%)	39.9	18.7	27.2	19.7%
M (%)	2.7	4.9	3.6	69.6%

(b) Example utterances by type

Type	Example
H	Do you want to watch a movie?
C-noncommand	How is the weather today?
C-command	Scroll down, Go back.
M	Show me sandwich shops. Oh, are you vegetarian?

Table 2: Out-of-domain corpora. "Single-user CB" is a corpus collected in same environment as the H-H-C in-domain data, except that only a single user was present.

Corpus	Addressee	Size
Single-user CB	H-C	21.9k words
Bing anchor text	H-C	1.3B bigrams
Fisher	H-H	21M words
ICSI meetings	H-H	0.7M words

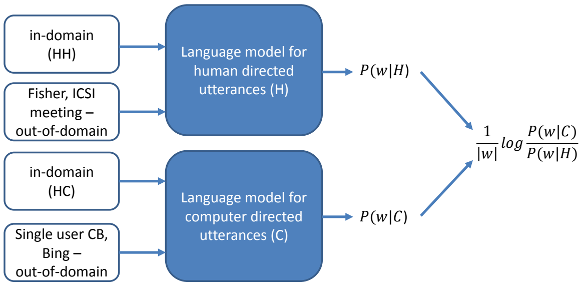
in-domain (HH)
Fisher, ICSI meeting – out-of-domain
in-domain (HC)
Single user CB, Bing – out-of-domain
Language model for human directed utterances (H)
Language model for computer directed utterances (C)
$P(w|H)$
$P(w|C)$
$\frac{1}{|w|} log \frac{P(w|C)}{P(w|H)}$

Figure 2: Language model-based score computation for addressee detection

tion of human- and computer-directed speech. The sizes and distribution of all utterance types, as well as sample utterances are shown in Table 1.

The ASR system used in the system was based on off-the-shelf acoustic models and had only the language model adapted to the domain, using very limited data. Consequently, as shown in the right-most column of Table 1(a), the word error rates (WERs) are quite high, especially for human-directed utterances. While these could be improved with targeted effort, we consider this a realistic application scenario, where in-domain training data is typically scarce, at least early in the development process. Therefore, any lexically based AD methods need to be robust to poor ASR accuracy.

2.2 Out-of-domain data

To replace the hard-to-obtain in-domain H-H-C data for training, we use the four out-of-domain corpora (two H-C and two H-H) shown in Table 2.

Single-user CB data comes from the same Conversational Browser system as the in-domain data, but with only one user present. This data can therefore be used for modeling H-C speech. Bing anchor text (Huang et al., 2010) is a large n-gram corpus of anchor text associated with links on web pages encountered by the Bing search engine. When users want to follow a link displayed on screen, they usually speak a variant of the anchor text for the link. We hypothesized that this corpus might aid the modeling of computer-noncommand type utterances in which such "verbal clicks" are frequent. Fisher telephone conversations and ICSI meetings are both corpora of human-directed speech. The Fisher corpus (Cieri et al., 2004) comprises two-person telephone conversations between strangers on prescribed topics. The ICSI meeting corpus (Janin et al., 2003) contains multiparty face-to-face technical discussions among colleagues.

3 Method

3.1 Language modeling for addressee detection

We use a lexical AD system that is based on modeling word n-grams in the two addressee-based utterance classes, H (for H-H) and C (for H-C utterances). This approach is similar to language model-based approaches to speaker and language recognition, and was shown to be quite effective for this task (Shriberg et al., 2012). Instead of making hard decisions, the system outputs a score that is the length-normalized likelihood ratio of the two classes:

$$\frac{1}{|w|} \log \frac{P(w|C)}{P(w|H)}, \tag{1}$$

where $|w|$ is the number of words in the recognition output w for an utterance. $P(w|C)$ and $P(w|H)$ are obtained from class-specific language models. Figure 2 gives a flow-chart of the score computation.

Class likelihoods are obtained from standard trigram backoff language models, using Witten-Bell discounting for smoothing (Witten and Bell, 1991). For combining various training data sources, we use language model adaptation by interpolation (Bellegarda, 2004). First, a separate model is trained from each source. The probability estimates from in-domain and out-of-domain models are then averaged in a weighted fashion:

$$P(w_k|h_k) = \lambda P_{in}(w_k|h_k) + (1-\lambda)P_{out}(w_k|h_k) \tag{2}$$

where w_k is the k-th word, h_k is the $(n-1)$-gram history for the word w_k. λ is the interpolation weight and is obtained by tuning a task-related metric on the development set. We investigated optimizing λ for either model perplexity or classification accuracy, as discussed below.

3.2 Part-of-speech-based modeling

So far we have only been modeling the lexical forms of words in utterances. If we encounter a word never before seen, it would appear as an out-of-vocabulary item in all class-specific language models, and not contribute much to the decision. More generally, if a word is rare, its n-gram statistics will be unreliable and poorly modeled by the system. (The sparseness issue is exacerbated by small amounts of training data as in our scenario.)

One common approach to deal with data sparseness in language modeling is to model n-grams over word classes rather than raw words (Brown et al., 1992). For example, if we have an utterance *How is the weather in Paris?*, the addressee probabilities are likely to be similar had we seen *London* instead of *Paris*. Therefore, replacing words with properly chosen word class labels can give better generalization from the observed training data. Among the many methods proposed to class words for language modeling purposes we chose part-of-speech (POS) tagging over other, purely data-derived classing algorithms (Brown et al., 1992), for two reasons. First, our goal here is not to minimize the perplexity of the data, but to enhance discrimination among utterance classes. Second, a data-driven class inference algorithm would suffer from the same sparseness issues when it comes to unseen and rare words (as no robust statistics are available to infer an unseen word's best class in the class induction step). A POS tagger, on the other hand, can do quite well on unseen words, using context and morphological cues.

A hidden Markov model tagger using POS-trigram statistics and context-independent class membership probabilities was used for tagging all LM training data. The tagger itself had been trained on the Switchboard (conversational telephone speech) transcripts of the Penn Treebank-3 corpus (Marcus et al., 1999), and used the 39 Treebank POS labels. To strike a compromise between generalization and discriminative power in the language model, we retained the top N most frequent word types from the in-domain training data as distinct tokens, and varied N as a metaparameter. Barzilay and Lee (2003) used a similar idea to generalize patterns by substituting words with slots. This strategy will tend to preserve words that are either generally frequent function and domain-independent words, capturing stylistic and syntactic patterns, or which are frequent domain-specific words, and can thus help characterize computer-directed utterances.

Here is a sample sentence and its transformed version:

> Original: *Let's find an Italian restaurant around this area.*
> POS-tagged: *Let's find an JJ NN around this area.*

The words except *Italian* and *restaurant* are unchanged because they are in the list of N most frequent words. We transformed all training and test data in this fashion and then modeled n-gram statistics as before. The one exception was the Bing anchor-text data, which was only available in the form of word n-grams (the sentence context required for accurate POS tagging was missing).

Table 3: Addressee detection performance (EER) with different training sets

	ASR	Transcript
Baseline (in-domain only)	31.1	17.3
Fisher+ICSI, Single-user CB+Bing (out-of-domain only)	27.8	14.2
Baseline + Fisher+ICSI, Single CB + Bing (both-all)	26.9	14.0
Baseline + ICSI, Single-user CB (both-small)	26.6	13.0

3.3 Evaluation metrics

Typically, an application-dependent threshold would be applied to the decision score to convert it into a binary decision. The optimal threshold is a function of prior class probabilities and error costs. As in Shriberg et al. (2012), we used equal error rate (EER) to compare systems, since we are interested in the discriminative power of the decision score independent of priors and costs. EER is the probability of false detections and misses at the operating point at which the two types of errors are equally probable. A prior-free metric such as EER is more meaningful than classification accuracy because the utterance type distribution is heavily skewed (Table 1), and because the rate of human- versus computer-directed speech can vary widely depending on the particular people, domain, and context. We also use classification accuracy (based on data priors) in one analysis below, because EERs are not comparable for different test data subdivisions.

3.4 Online model

The actual dialog system used in this work processes utterances after receiving an entire segment of speech from the recognition subsystem. However, we envision that a future version of the system would perform addressee detection in an online manner, making a decision as soon as enough evidence is gathered. This raises the question how soon the addressee can be detected once the user starts speaking. We simulate this processing mode using a windowed AD model.

As shown in Figure 3, we define windows starting at the beginning of the utterance and investigate how AD performance changes as a function of window size. We use only the words and n-grams falling completely within a given window. For example, the word *find* would be excluded from Window 1 in Figure 3.

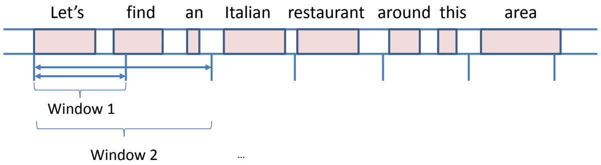

Figure 3: The window model

The benefit of early detection in this case is that once speech is classified as human-directed, it does not need to be sent to the speech recognizer and subsequent semantic processing. This saves processing time, especially if processing happens on a server. Based on the window model performance, we can assess the feasibility of an online AD model, which can be approached by shifting the detection window through time and finding addressee changes.

4 Results and Discussion

Table 3 compares the performance of our system using various training data sources. For diagnostic purposes we also compare performance based on recognized words (the realistic scenario) to that based on human transcripts (idealized, best-case word recognition).

Somewhat surprisingly, the system trained on out-of-domain data alone performs better by 3.3 EER points on ASR output and 3.1 points on transcripts compared to the in-domain baseline. Combining in-domain and out-of-domain data (both-all, both-small) gives about 1 point additional EER gain. Note that training on in-domain data plus the smaller-size out-of-domain corpora (both-small) is better than using all available data (both-all).

Figure 4 shows the detection error trade-off (DET) between false alarm and miss errors for the

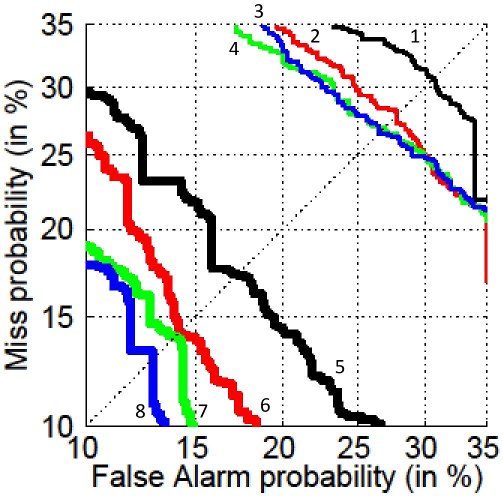

Figure 4: Detection error trade-off (DET) curves for the systems in Table 3. Thin lines at the top right corner use ASR output (1-4); thick lines at the bottom left corner use reference transcripts (5-8). Each line number represents one of the systems in Table 3: 1,5 = in-domain only, 2,6 = out-of-domain only, 4,7 = both-all, 3,8 = both-small.

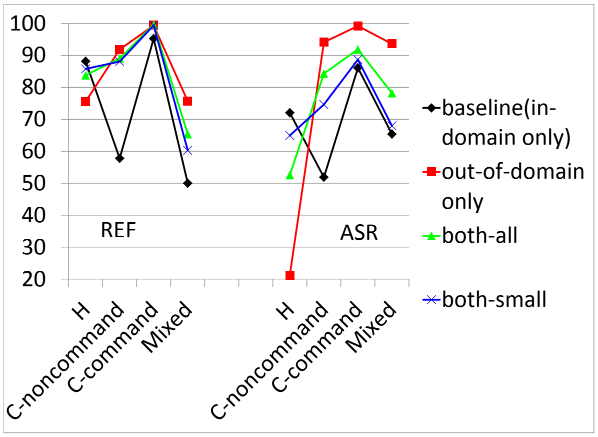

Figure 5: AD accuracies by utterance type

Table 4: Perplexities (computed on dev set ASR words) by utterance type, for different training corpora. Interpolation refers to the combination of the three models listed in each case.

	Test class	
Training set	H-C	H-H
In-domain H-C (ASR)	257	1856
Single-user CB	104	1237
Bing anchor text	356	789
Interpolation	58	370
In-domain H-H (ASR)	887	1483
Fisher	995	795
ICSI meeting	2007	1583
Interpolation	355	442

systems in Table 3. The DET plot depicts performance not only at the EER operating point (which lies on the diagonal), but over the range of possible trade-offs between false alarm and miss error rates. As can be seen, replacing or combining in-domain data with out-of-domain data gives clear performance gains, regardless of operating point (score threshold), and for both reference and recognized words.

Figure 5 shows H-H vs. H-C classification accuracies on each of the four utterance subtypes listed in Table 1. It is clear that computer-command utterances are the easiest to classify; the accuracy is more than 90% using transcripts, and more than 85% using ASR output. This is not surprising, since commands are from a fixed small set of phrases. The biggest gain from use of out-of-domain data is found for computer-directed noncommand utterances. This is helpful, since in general it is the noncommand computer-directed utterances (rather than the commands) that are highly confusable with human-directed utterances: both use unconstrained natural language. We note that H-H utterance are very poorly recognized in the ASR condition when only out-of-domain data is used. This may be because the human-human corpora used in training consist of transcripts, whereas the ASR output for human-directed utterances is very errorful, creating a severe train-test mismatch.

As for the optimization of the mixing weight λ, we found that minimizing perplexity on the development set of each class is effective. This is a standard optimization approach for interpolated language models, and can be carried out efficiently using an expectation maximization algorithm. We also tried search-based optimization using the classification metric (EER) as the criterion. While this approach could theoretically give better results (since perplexity is not a discriminative criterion) we found no significant improvement in our experiments.

Table 4 shows the perplexities by class of language models trained on different corpora. We can take these as an indication of training/test mismatch (lower perplexity indicating better match). We also find substantial perplexity reductions from interpolating models. In order to make perplexities comparable, we trained all models using the union of the vocabularies from the different sources.

In spite of perplexity being a good way to optimize the *weighting* of sources, it is not clear that it is a good criterion for *selecting* data sources. For example, we see that the Fisher model has a much lower perplexity on H-H utterances than the ICSI meeting model. However, as reflected in Table 3, the H language model that leaves out the Fisher data actually performed better. The most likely explanation is that the Fisher corpus is an order of magnitude larger than the ICSI corpus, and that sheer data size, not stylistic similarity, may account for the lower perplexity of the Fisher model. Further investigation is needed regarding good criteria for corpus selection for classification tasks such as AD.

Table 5 shows the EER performance of the POS-based model, for various sizes N of the most-frequent word list. We observe that the partial replacement of words with POS tags indeed improves over the baseline model performance, by 1.5 points on ASR output and by 1.1 points on transcripts. We also see that the gain over the corresponding word-only model is largest for the in-domain baseline model, and less or non-existent for the out-of-domain model. This is consistent with the notion that the in-domain model suffers the most from data sparseness, and therefore has the most to gain from better generalization.

Interpolating with out-of-domain data still helps here. The optimal N differs for ASR output versus transcripts. The POS-based model with $N = 300$ improves the EER by 0.5 points on ASR output, and $N = 1000$ improves the EER by 0.8 points on transcripts. Here we use relatively large amounts of training data, thus the performance gain is smaller, though still meaningful.

Figure 6 shows the performance of the system using time windows anchored at the beginnings of utterances. We incrementally increase the window width from 0.5 seconds to 3 seconds and compare results to using full utterances. The leveling off of the error plots indicates that most addressee information is contained in the first 1 to 1.5 seconds, although some additional information is found in the later part of utterances (the plots never level off completely). This pattern holds for both in-domain and out-of-domain training, as well as for combined models.

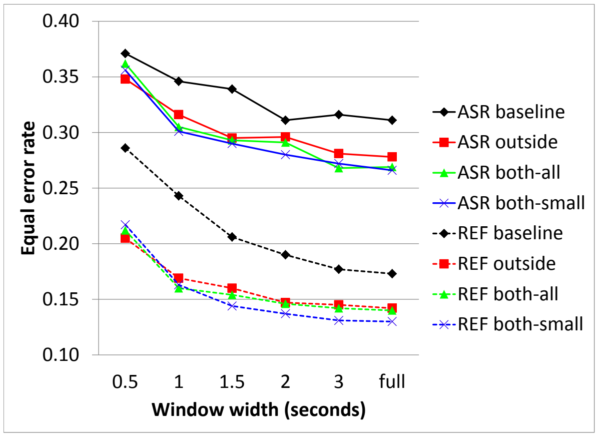

Figure 6: Simulated online performance on incremental windows

Table 6: The top 15 first words in utterances

ASR H-C	Transcript H-C	ASR H-H	Transcript H-H
go	go	play	I
scroll	scroll	go	ohh
start	start	is	so
show	stop	it	yeah
stop	show	what	it’s
bing	find	this	you
search	Bing	show	uh
find	search	how	okay
play	pause	bing	what
pause	play	select	it
look	look	okay	and
what	uh	does	that’s
select	what	start	is
how	how	so	no
the	ohh	I	we

To give an intuitive understanding of where this early addressee-relevant information comes from, we tabulated the top 15 word unigrams in each utterance class, are shown in Table 6. Note that the substantial differences between the third and fourth columns in the table reflect the high ASR error rate for human-directed utterances, whereas

Table 5: Performance of POS-based model with various top-N word lists (EER)

	Training data	top100	top200	top300	top400	top500	top1000	top2000	Original
ASR	baseline	31.6	31.0	**29.6**	30.1	30.2	31.4	31.5	31.1
	out-of-domain only	36.5	37.0	37.2	36.9	36.8	36.6	37.3	**27.8**
	both-all	28.2	26.6	**26.1**	26.7	27.4	26.9	27.6	26.9
	both-small	28.0	26.5	**26.2**	26.6	26.4	26.3	26.5	26.6
REF	baseline	17.1	**16.2**	16.6	17.1	16.7	17.0	17.2	17.3
	out-of-domain only	17.6	17.6	17.5	17.2	17.1	17.2	18.1	**14.2**
	both-all	**12.5**	**12.5**	**12.5**	12.7	12.8	13.2	13.5	14.0
	both-small	13.0	13.2	12.8	13.2	12.8	**12.2**	12.7	13.0

for computer-directed utterances, the frequent first words are mostly recognized correctly.

In computer-directed utterances we see mostly command verbs, which, due to the imperative syntax of these commands occur in utterance-initial position. Human-directed utterances are characterized by subject pronouns such as *I* and *it*, or answer particles such as *yeah* and *okay*, which likewise occur in initial position. Based on word frequency and syntax alone it is thus clear why the beginnings of utterances contain strong lexical cues.

5 Conclusion

We explored the use of outside data for training lexical addressee detection systems for the human-human-computer scenario. Advantages include saving the time and expense of an in-domain data collection, as well as performance gains even when some in-domain data is available. We show that H-C training data can be obtained from a single-user H-C collection, and that H-H speech can be modeled using general conversational speech. Using the outside training data, we obtain results that are even better than results using matched (but smaller) H-H-C training data. Results can be improved considerably by adapting H-C and H-H language models with small amounts of matched H-H-C data, via interpolation. The main reason for the improvement is better detection of computer-directed noncommand utterances, which tend to be confusable with human-directed utterances. Another effective way to overcome scarce training data is to replace the less frequent words with part-of-speech labels. In both baseline and interpolated model, we found that POS-based models that keep an appropriate number of the top N most frequent word types can further improve the system's performance.

In a second study we found that the most salient phrases for lexical addressee detection occur within the first 1 to 1.5 seconds of speech in each utterance. It reflects a syntactic tendency of class-specific words to occur utterance-initially, which shows the feasibility of the online AD system.

Acknowledgments

We thank our Microsoft colleagues Madhu Chinthakunta, Dilek Hakkani-Tür, Larry Heck, Lisa Stiefelman, and Gokhan Tür for developing the dialog system used in this work, as well as for many valuable discussions. Ashley Fidler was in charge of much of the data collection and annotation required for this study. We also thank Dan Jurafsky for useful feedback.

References

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings HLT-NAACL 2003*, pages 16–23, Edmonton, Canada.

Jerome R. Bellegarda. 2004. Statistical language model adaptation: review and perspectives. *Speech Communication*, 42:93–108.

Dan Bohus and Eric Horvitz. 2011. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings ACL SIGDIAL*, pages 98–109, Portland, OR.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings 4th International Conference on Language Resources and Evaluation*, pages 69–71, Lisbon.

John Dowding, Richard Alena, William J. Clancey, Maarten Sierhuis, and Jeffrey Graham. 2006. Are you talking to me? dialogue systems supporting mixed teams of humans and robots. In *Proccedings AAAI Fall Symposium: Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic Systems*, Washington, DC.

Dilek Hakkani-Tür, Gokhan Tur, and Larry Heck. 2011a. Research challenges and opportunities in mobile applications [dsp education]. *IEEE Signal Processing Magazine*, 28(4):108 –110.

Dilek Z. Hakkani-Tür, Gökhan Tür, Larry P. Heck, and Elizabeth Shriberg. 2011b. Bootstrapping domain detection using query click logs for new domains. In *Proceedings Interspeech*, pages 709–712.

Jian Huang, Jianfeng Gao, Jiangbo Miao, Xiaolong Li, Kuansang Wang, and Fritz Behr. 2010. Exploring web scale language models for search query processing. In *Proceedings 19th International Conference on World Wide Web*, pages 451–460, Raleigh, NC.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *Proceedings IEEE ICASSP*, volume 1, pages 364–367, Hong Kong.

Michael Katzenmaier, Rainer Stiefelhagen, and Tanja Schultz. 2004. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings 6th International Conference on Multimodal Interfaces*, ICMI, pages 144–151, New York, NY, USA. ACM.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3. Linguistic Data Consortium, catalog item LDC99T42.

Rieks op den Akker and David Traum. 2009. A comparison of addressee detection methods for multiparty conversations. In *Proceedings of Diaholmia*, pages 99–106.

Tim Paek, Eric Horvitz, and Eric Ringger. 2000. Continuous listening for unconstrained spoken dialog. In *Proceedings ICSLP*, volume 1, pages 138–141, Beijing.

Daniel Reich, Felix Putze, Dominic Heger, Joris Ijsselmuiden, Rainer Stiefelhagen, and Tanja Schultz. 2011. A real-time speech command detector for a smart control room. In *Proceedings Interspeech*, pages 2641–2644, Florence.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Larry Heck. 2012. Learning when to listen: Detecting system-addressed speech in human-human-computer dialog. In *Proceedings Interspeech*, Portland, OR.

Ian H. Witten and Timothy C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094.