Robustness Issues in a Data-Driven Spoken Language Understanding System

Yulan He and **Steve Young**
Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, England
{yh213, sjy}@eng.cam.ac.uk

Abstract

Robustness is a key requirement in spoken language understanding (SLU) systems. Human speech is often ungrammatical and ill-formed, and there will frequently be a mismatch between training and test data. This paper discusses robustness and adaptation issues in a statistically-based SLU system which is entirely data-driven. To test robustness, the system has been tested on data from the Air Travel Information Service (ATIS) domain which has been artificially corrupted with varying levels of additive noise. Although the speech recognition performance degraded steadily, the system did not fail catastrophically. Indeed, the rate at which the end-to-end performance of the complete system degraded was significantly slower than that of the actual recognition component. In a second set of experiments, the ability to rapidly adapt the core understanding component of the system to a different application within the same broad domain has been tested. Using only a small amount of training data, experiments have shown that a semantic parser based on the Hidden Vector State (HVS) model originally trained on the ATIS corpus can be straightforwardly adapted to the somewhat different DARPA Communicator task using standard adaptation algorithms. The paper concludes by suggesting that the results presented provide initial support to the claim that an SLU system which is statistically-based and trained entirely from data is intrinsically robust and can be readily adapted to new applications.

1 Introduction

Spoken language is highly variable as different people use different words and sentence structures to convey the same meaning. Also, many utterances are grammatically-incorrect or ill-formed. It thus remains an open issue as to how to provide robustness for large populations of non-expert users in spoken dialogue systems. The key component of a spoken language understanding (SLU) system is the semantic parser, which translates the users' utterances into semantic representations. Traditionally, most semantic parser systems have been built using hand-crafted semantic grammar rules and so-called *robust parsing* (Ward and Issar, 1996; Seneff, 1992; Dowding et al., 1994) is used to handle the ill-formed user input in which word patterns corresponding to semantic tokens are used to fill slots in different semantic frames in parallel. The frame with the highest score then yields the selected semantic representation.

Formally speaking, the robustness of language (recognition, parsing, etc.) is a measure of the ability of human speakers to communicate despite incomplete information, ambiguity, and the constant element of surprise (Briscoe, 1996). In this paper, two aspects of SLU system performance are investigated: noise robustness and adaptability to different applications. For the former, we expect that an SLU system should maintain acceptable performance when given noisy input speech data. This requires, the understanding components of the SLU system to be able to correctly interpret the meaning of an utterance even when faced with recognition errors. For the latter, the SLU system should be readily adaptable to a different application using a relatively small set (e.g. less than 100) of adaptation utterances.

The rest of the paper is organized as follows. An overview of our data-driven SLU system is outlined in section 2. Experimental results on performance under a range of SNRs are then presented in section 3. Section 4 discusses adaptation of the HVS model to new applica-

tions. Finally, section 5 concludes the paper.

2 System Overview

Spoken language understanding (SLU) aims to interpret the meanings of users' utterances and respond reasonably to what users have said. A typical architecture of an SLU system is given in Fig. 1, which consists of a speech recognizer, a semantic parser, and a dialog act decoder. Within a statistical framework, the SLU problem can be factored into three stages. First the speech recognizer recognizes the underlying word string W from each input acoustic signal A, i.e.

$$\hat{W} = \underset{W}{\text{argmax}}\, P(W|A) = \underset{W}{\text{argmax}}\, P(A|W)P(W) \quad (1)$$

then the semantic parser maps the recognized word string $\hat{W}$ into a set of semantic concepts C

$$\hat{C} = \underset{C}{\text{argmax}}\, P(C|\hat{W}) \quad (2)$$

and finally the dialogue act decoder infers the user's dialog acts or goals by solving

$$\hat{G_u} = \underset{G_u}{\text{argmax}}\, P(G_u|\hat{C}) \quad (3)$$

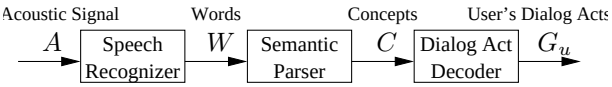


Figure 1: Typical structure of a spoken language understanding system.

The sequential decoding described above is suboptimal since the solution at each stage depends on an exact solution to the previous stage. To reduce the effect of this approximation, a word lattice or N-best word hypotheses can be retained instead of the single best string $\hat{W}$ as the output of the speech recognizer. The semantic parse results may then be incorporated with the output from the speech recognizer to rescore the N-best list as below.

$$\begin{aligned} \hat{C}, \hat{W} &\approx \underset{C,W \in L_N}{\text{argmax}}\, P(A|W)P(W)P(C|W) \\ &\approx \underset{C,W \in L_N}{\text{argmax}}\, P(A|W)P(W)^{\gamma}P(C|W)^{\alpha} \quad (4) \end{aligned}$$

where $P(A|W)$ is the acoustic probability from the first pass, $P(W)$ is the language modelling likelihood, $P(C|W)$ is the semantic parse score, L_N denotes the N-best list, α is a semantic parse scale factor, and γ is a grammar scale factor.

In the system described in this paper, each of these stages is modelled separately. We use a standard HTK-based (HTK, 2004) Hidden Markov Model (HMM) recognizer for recognition, the Hidden Vector State (HVS) model for semantic parsing (He and Young, 2003b), and Tree-Augmented Naive Bayes networks (TAN) (Friedman et al., 1997) for dialog act decoding.

The speech recognizer comprises 14 mixture Gaussian HMM state-clustered cross-word triphones augmented by using heteroscedastic linear discriminant analysis (HLDA) (Kumar, 1997). Incremental speaker adaptation based on the maximum likelihood linear regression (MLLR) method (Gales and Woodland, 1996) was performed during the test with updating being performed in batches of five utterances per speaker.

The Hidden Vector State (HVS) model (He and Young, 2003b) is a hierarchical semantic parser which associates each state of a push-down automata with the state of a HMM. State transitions are factored into separate stack pop and push operations and then constrained to give a tractable search space. The result is a model which is complex enough to capture hierarchical structure but which can be trained automatically from unannotated data.

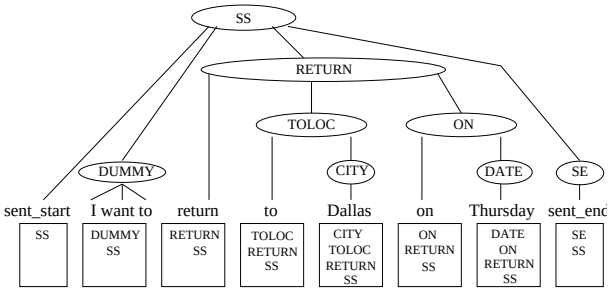


Figure 2: Example of a parse tree and its vector state equivalent.

Let each state at time t be denoted by a vector of D_t semantic concept labels (tags) $\mathbf{c}_t = [c_t[1], c_t[2], ..c_t[D_t]]$ where $c_t[1]$ is the preterminal concept and $c_t[D_t]$ is the root concept (SS in Figure 2). Given a word sequence W, concept vector sequence $\mathbf{C}$ and a sequence of stack pop operations N, the joint probability of $P(W, \mathbf{C}, N)$ can be decomposed as

$$\begin{aligned} P(W, \mathbf{C}, N) = \prod_{t=1}^{T} & P(n_t|\mathbf{c}_{t-1}) \cdot \\ & P(c_t[1]|c_t[2 \cdots D_t]) \cdot P(w_t|\mathbf{c}_t) \quad (5) \end{aligned}$$

where $\mathbf{c}_t$ at word position t is a vector of D_t semantic concept labels (tags), n_t is the vector stack shift operation and takes values in the range $0, \cdots, D_{t-1}$ where D_{t-1} is the stack size at word position $t-1$, and $c_t[1] = c_{w_t}$ is the new preterminal semantic tag assigned to word w_t at word position t.

Thus, the HVS model consists of three types of probabilistic move:

1. popping semantic tags off the stack;

2. pushing a pre-terminal semantic tag onto the stack;

3. generating the next word.

The dialog act decoder was implemented using the Tree-Augmented Naive Bayes (TAN) algorithm (Friedman et al., 1997), which is an extension of Naive Bayes Networks. One TAN was used for each dialogue act or goal G_u, the semantic concepts C_i which serve as input to its corresponding TAN were selected based on the mutual information (MI) between the goal and the concept. Naive Bayes networks assume all the concepts are conditionally independent given the value of the goal. TAN networks relax this independence assumption by adding dependencies between concepts based on the conditional mutual information (CMI) between concepts given the goal. The goal prior probability $P(G_u)$ and the conditional probability of each semantic concept C_i given the goal G_u, $P(C_i|G_u)$ are learned from the training data. Dialogue act detection is done by picking the goal with the highest posterior probability of G_u given the particular instance of concepts $C_1 \cdots C_n$, $P(G_u|C_1 \cdots C_n)$.

3 Noise Robustness

The ATIS corpus which contains air travel information data (Dahl et al., 1994) has been chosen for the SLU system development and evaluation. ATIS was developed in the DARPA sponsored spoken language understanding programme conducted from 1990 to 1995 and it provides a convenient and well-documented standard for measuring the end-to-end performance of an SLU system. However, since the ATIS corpus contains only clean speech, corrupted test data has been generated by adding samples of background noise to the clean test data at the waveform level.

3.1 Experimental Setup

The experimental setup used to evaluate the SLU system was similar to that described in (He and Young, 2003a). As mentioned in section 2, the SLU system consists of three main components, a standard HTK-based HMM recognizer, the HVS semantic parser, and the TAN dialogue act (DA) decoder. Each of the three major components are trained separately. The acoustic speech signal in the ATIS training data is modelled by extracting 39 features every 10ms: 12 cepstra, energy, and their first and second derivatives. This data is then used to train the speaker-independent, continuous speech recognizer. The HVS semantic parser is trained on the unannotated utterances using EM constrained by the domain-specific lexical class information and the dominance relations built into the abstract annotations (He and Young, 2003b). In the case of ATIS, the lexical classes can be extracted automatically from the relational database, whilst abstract semantic annotations for each utterance are automatically derived from the accompanying SQL queries of the training utterances. The dialogue act decoder is trained using the main topics or goals and the key semantic concepts extracted automatically from the reference SQL queries

Performance is measured at both the component and the system level. For the former, the recognizer is evaluated by word error rate, the parser by concept slot retrieval rate using an F-measure metric (Goel and Byrne, 1999), and the dialog act decoder by detection rate. The overall system performance is measured using the standard NIST "query answer" rate.

In the expriments reported here, car noise from the NOISEX-92 (Varga et al., 1992) database was added to the ATIS-3 NOV93 and DEC94 test sets. In order to obtain different SNRs, the noise was scaled accordingly before adding to the speech signal.

3.2 Experimental Results

Robust spoken language understanding components should be able to compensate for the weakness of the speech recognizer. That is, ideally they should be capable of generating the correct meaning of an utterance even if it is recognized wrongly by a speech recognizer. At minimum, the performance of the understanding components should degrade gracefully as recognition accuracy degrades.

Figure 3 gives the system performance on the corrupted test data with additive noise ranging from 25dB to 10dB SNR. The label "clean" in the X-axis denotes the original clean speech data without additive noise. Note that the recognition results on the corrupted test data were obtained directly using the original clean speech HMM models without retraining for the noisy conditions. The upper portion of Figure 3 shows the end-to-end performance in terms of query answer error rate for the NOV93 and DEC94 test sets. For easy reference, WER is also shown. The individual component performance, F-measure for the HVS semantic parser and dialogue act (DA) detection accuracy for the DA decoder, are illustrated in the lower portion of Figure 3. For each test set, the performance on the rescored word hypotheses is given as well. This incorporates the semantic parse scores into the acoustic and language modelling likelihoods to rescore the 25-best word lists from the speech recognizer.

It can be observed that the system gives fairly stable performance at high SNRs and then the recognition accuracy degrades rapidly in the presence of increasing noise. At 20dB SNR, the WER for the NOV93 test set increases by 1.6 times relative to clean whilst the query answer error rate increases by only 1.3 times. On decreasing

the SNR to 15dB, the system performance degrades significantly. The WER increases by 3.1 times relative to clean but the query answer error rate increases by only 1.7 times. Similar figures were obtained for the DEC94 test set.

The above suggests that the end-to-end performance measured in terms of answer error rate degrades more slowly compared to the recognizer WER as the noise level increases. This demonstrates that the statistically-based understanding components of the SLU system, the semantic parser and the dialogue act decoder, are relatively robust to degrading recognition performance.

Regarding the individual component performance, the dialogue act detection accuracy appears to be less sensitive to decreasing SNR. This is probably a consequence of the fact that the Bayesian networks are set up to respond to only the presence or absence of semantic concepts or slots, regardless of the actual values assigned to them. In another words, the performance of the dialogue act decoder is not affected by the mis-recognition of individual words, but only by a failure to detect the presence of a semantic concept. It can also be observed from Figure 3 that the F-measure needs to be better than 85% in order to achieve acceptable end-to-end performance.

4 Adaptation to New Applications

Statistical model adaptation techniques are widely used to reduce the mismatch between training and test or to adapt a well-trained model to a novel domain. Commonly used techniques can be classified into two categories, Bayesian adaptation which uses a maximum *a posteriori* (MAP) probability criteria (Gauvain and Lee, 1994) and transformation-based approaches such as maximum likelihood linear regression (MLLR) (Gales and Woodland, 1996), which uses a maximum likelihood (ML) criteria. In recent years, MAP adaptation has been successfully applied to n-gram language models (Bacchiani and Roark, 2003) and lexicalized PCFG models (Roark and Bacchiani, 2003). Luo *et al.* have proposed transformation-based approaches based on the Markov transform (Luo et al., 1999) and the Householder transform (Luo, 2000), to adapt statistical parsers. However, the optimisation processes for the latter are complex and it is not clear how general they are.

Since MAP adaptation is straightforward and has been applied successfully to PCFG parsers, it has been selected for investigation in this paper. Since one of the special forms of MAP adaptation is interpolation between the in-domain and out-of-domain models, it is natural to also consider the use of non-linear interpolation and hence this has been studied as well [1].

[1]Experiments using linear interpolation have also been conducted but it was found that the results are worse than those obtained using MAP adaptation or log-linear interpolation.

4.1 MAP Adaptation

Bayesian adaptation reestimates model parameters directly using adaptation data. It can be implemented via maximum *a posteriori* (MAP) estimation. Assuming that model parameters are denoted by Θ, then given observation samples Y, the MAP estimate is obtained as

$$\Theta_{MAP} = \underset{\Theta}{\operatorname{argmax}} P(\Theta|Y) = \underset{\Theta}{\operatorname{argmax}} P(Y|\Theta)P(\Theta) \tag{6}$$

where $P(Y|\Theta)$ is the likelihood of the adaptation data Y and model parameters Θ are random vectors described by their probabilistic mass function (pmf) $P(\Theta)$, also called the prior distribution.

In the case of HVS model adaptation, the objective is to estimate probabilities of discrete distributions over vector state stack shift operations and output word generation. Assuming that they can be modelled under the multinomial distribution, for mathematical tractability, the conjugate prior, the Dirichlet density, is normally used. Assume a parser model $P(W, C)$ for a word sequence W and semantic concept sequence C exists with J component distributions P_j each of dimension K, then given some adaptation data W_l, the MAP estimate of the kth component of P_j, $\hat{P}_j(k)$, is

$$\hat{P}_j(k) = \frac{\sigma_j}{\sigma_j + \tau}\tilde{P}_j(k) + \frac{\tau}{\sigma_j + \tau}P_j(k) \tag{7}$$

where $\sigma_j = \sum_{k=1}^{K} \sigma_j(k)$ in which $\sigma_j(k)$ is defined as the total count of the events associated with the kth component of P_j summed across the decoding of all adaptation utterances W_l, τ is the prior weighting parameter, $P_j(k)$ is the probability of the original unadapted model, and $\tilde{P}_j(k)$ is the empirical distribution of the adaptation data, which is defined as

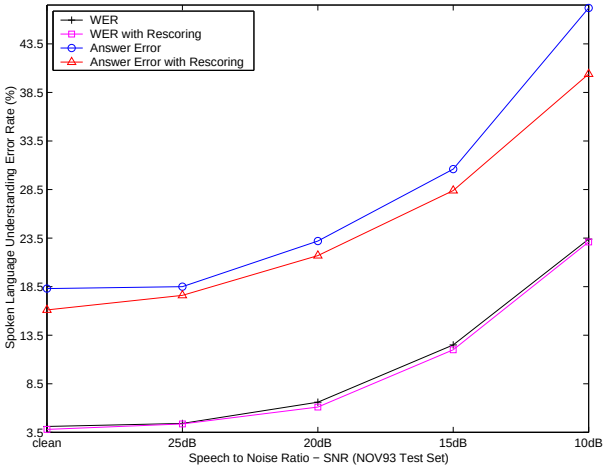
$$\tilde{P}_j(k) = \frac{\sigma_j(k)}{\sum_{i=1}^{K} \sigma_j(i)} \tag{8}$$

As discussed in section 2, the HVS model consists of three types of probabilistic move. The MAP adaptation technique can be applied to the HVS model by adapting each of these three component distributions individually.

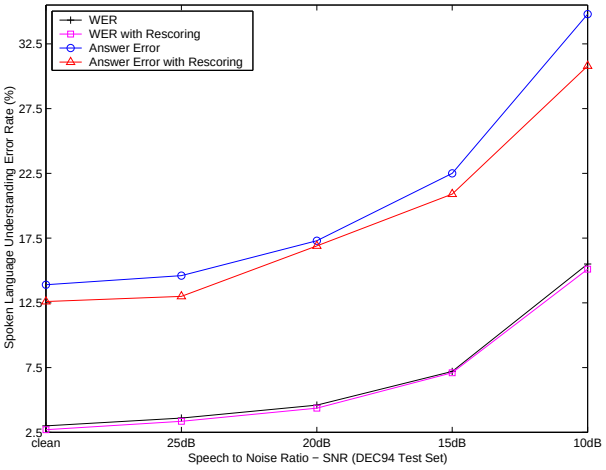
4.2 Log-Linear Interpolation

Log-linear interpolation has been applied to language model adaptation and has been shown to be equivalent to a constrained minimum Kullback-Leibler distance optimisation problem(Klakow, 1998).

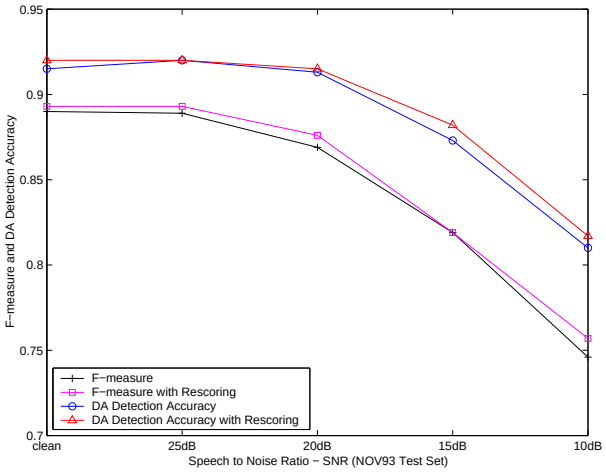
Following the notation introduced in section 4.1, where $P_j(k)$ is the probability of the original unadapted model, and $\tilde{P}_j(k)$ is the empirical distribution of the adaptation



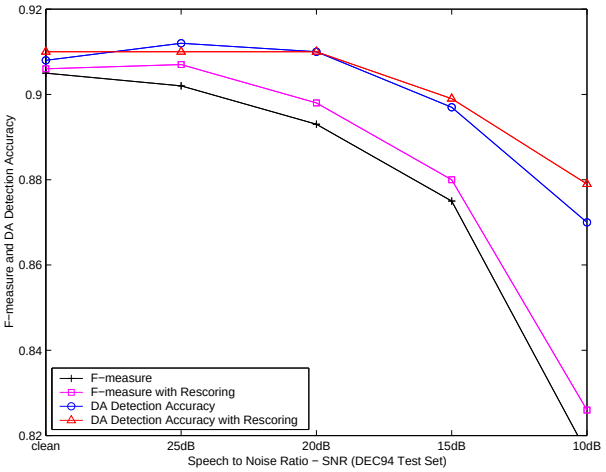
(a) NOV93 End-to-End Performance



(c) DEC94 End-to-End Performance



(b) NOV93 Component Performance



(d) DEC94 Component Performance

Figure 3: SLU system performance vs SNR.

data, denote the final adapted model probability as $\hat{P}_j(k)$. It is assumed that the Kullback-Leibler distance of the adapted model to the unadapted and empirically determined model is

$$D(\hat{P}_j(k) \parallel P_j(k)) = d_1 \quad (9)$$

$$D(\hat{P}_j(k) \parallel \tilde{P}_j(k)) = d_2 \quad (10)$$

Given an additional model probability $\bar{P}_j(k)$ whose distance to $\hat{P}_j(k)$ should be kept small, and introducing Lagrange multipliers $\lambda_1^{'}$ and $\lambda_2^{'}$ to ensure that constraints 9 and 10 are satistifed, yields

$$\mathcal{D} = D(\hat{P}_j(k) \parallel \bar{P}_j(k)) + \lambda_1^{'}(D(\hat{P}_j(k) \parallel P_j(k)) - d_1) + \lambda_2^{'}(D(\hat{P}_j(k) \parallel \tilde{P}_j(k)) - d_2) \quad (11)$$

Minimizing $\mathcal{D}$ with respect to $\hat{P}_j(k)$ yields the required distribution.

With some manipulation and redefinition of the Lagrange Multipliers, it can be shown that

$$\hat{P}_j(k) = \frac{1}{Z_\lambda} P_j(k)^{\lambda_1} \tilde{P}_j(k)^{\lambda_2} \quad (12)$$

where $\bar{P}_j(k)$ has been assumed to be a uniform distribution which is then absorbed into the normalization term Z_λ.

The computation of Z_λ is very expensive and can usually be dropped without significant loss in performance (Martin et al., 2000). For the other parameters, λ_1 and λ_2, the generalized iterative scaling algorithm or the simplex method can be employed to estimate their optimal settings.

4.3 Experiments

To test the portability of the statistical parser, the initial experiments reported here are focussed on assessing the adaptability of the HVS model when it is tested in a domain which covers broadly similar concepts, but comprises rather different speaking styles. To this end, the flight information subset of the DARPA Communicator Travel task has been used as the target domain (CUData, 2004). By limiting the test in this way, we ensure that the dimensionalities of the HVS model parameters remain the same and no new semantic concepts are introduced by the adaptation training data.

The baseline HVS parser was trained on the ATIS corpus using 4978 utterances selected from the context-independent (Class A) training data in the ATIS-2 and ATIS-3 corpora. The vocabulary size of the ATIS training corpus is 611 and there are altogether 110 semantic concepts defined. The parser model was then adapted using utterances relating to flight reservation from the DARPA Communicator data. Although the latter bears similarities to the ATIS data, it contains utterances of a different style and is often more complex. For example, Communicator contains utterances on multiple flight legs, information which is not available in ATIS.

To compare the adapted ATIS parser with an in-domain Communicator parser, a HVS model was trained from scratch using 10682 Communicator training utterances. The vocabulary size of the in-domain Communicator training data is 505 and a total of 99 semantic concepts have been defined. For all tests, a set of 1017 Communicator test utterances was used.

Table 1 lists the recall, precision, and F-measure results obtained when tested on the 1017 utterance DARPA Communicator test set. The baseline is the unadapted HVS parser trained on the ATIS corpus only. The in-domain results are obtained using the HVS parser trained solely on the 10682 DARPA training data. The other rows of the table give the parser performance using MAP and log-linear interpolation based adaptation of the baseline model using 50 randomly selected adaptation utterances.

System	*Recall*	*Precision*	*F-measure*
Baseline	79.81%	87.14%	83.31%
In-domain	87.18%	91.89%	89.47%
MAP	86.74%	91.07%	88.85%
Log-Linear	86.25%	92.35%	89.20%

Table 1: Performance comparison of adaptation using MAP or log-linear interpolation.

Since we do not yet have a reference database for the DARPA Communicator task, it is not possible to conduct the end-to-end performance evaluation as in section 3. However, the experimental results in section 3.2 indicate that the F-measure needs to exceed 85% to give acceptable end-to-end performance (see Figure 3). Therefore, it can be inferred from Table 1 that the unadapted ATIS parser model would perform very badly in the new Communicator application whereas the adapted models would give performance close to that of a fully trained in-domain model.

Figure 4 shows the parser performance versus the number of adaptation utterances used. It can be observed that when there are only a few adaptation utterances, MAP adaptation performs significantly better than log-linear interpolation. However above 25 adaptation utterances, the converse is true. The parser performance saturates when the number of adaptation utterances reaches 50 for both techniques and the best performance overall is given by the parser adapted using log-linear interpolation. The performance of both models however degrades when the number of adaptation utterances exceeds 100, possibly due to model overtraining. For this particular application, we conclude that just 50 adaptation utterances would be sufficient to adapt the baseline model to give comparable

results to the in-domain Communicator model.

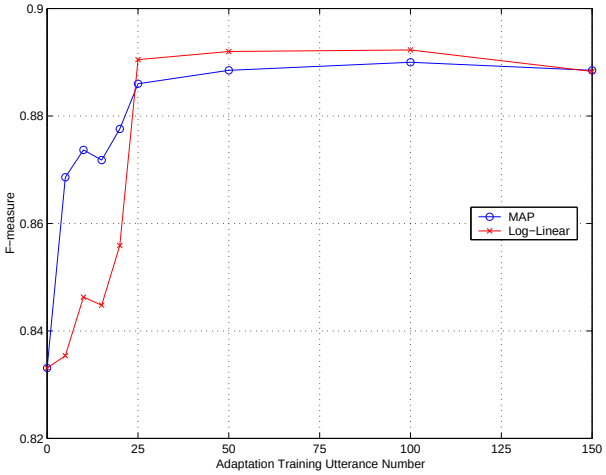


Figure 4: F-measure vs amount of adaptation training data.

5 Conclusions

The spoken language understanding (SLU) system discussed in this paper is entirely statistically based. The recogniser uses a HMM-based acoustic model and an n-gram language model, the semantic parser uses a hidden vector state model and the dialogue act decoder uses Bayesian networks. The system is trained entirely from data and there are no heuristic rules. One of the major claims motivating the design of this type of system is that its fully-statistical framework makes it intrinsically robust and readily adaptable to new applications. The aim of this paper has been to investigate this claim experimentally via two sets of experiments using a system trained on the ATIS corpus.

In the first set of experiments, the acoustic test data was corrupted with varying levels of additive car noise. The end-to-end system performance was then measured along with the individual component performances. It was found that although the addition of noise had a substantial effect on the word error rate, its relative influence on both the semantic parser slot/value retrieval rate and the dialogue act detection accuracy was somewhat less. Overall, the end-to-end error rate degraded relatively more slowly than word error rate and perhaps most importantly of all, there was no catastrophic failure point at which the system effectively stops working, a situation not uncommon in current rule-based systems.

In the second set of experiments, the ability of the semantic decoder component to be adapted to another application was investigated. In order, to limit the issues to parameter mismatch problems, the new application chosen (Communicator) covered essentially the same set of concepts but was a rather different corpus with different user speaking styles and different syntactic forms. Overall, we found that moving a system trained on ATIS to this new application resulted in a 6% absolute drop in F-measure on concept accuracy (i.e. a 62% relative increase in parser error) and by extrapolation with the results in the ATIS domain, we infer that this would make the non-adapted system essentially unusable in the new application. However, when adaptation was applied using only 50 adaptation sentences, the loss of concept accuracy was mostly restored. Specifically, using log-linear adaptation, the out-of-domain F-measure of 83.3% was restored to 89.2% which is close to the in-domain F-measure of 89.5%.

Although these tests are preliminary and are based on off-line corpora, the results do give positive support to the initial claim made for statistically-based spoken language systems, i.e. that they are robust and they are readily adaptable to new or changing applications.

Acknowledgements

The authors would like to thank Mark Gales for providing the software to generate the corrupted speech data with additive noise.

References

M. Bacchiani and B. Roark. 2003. Unsupervised language model adaptation. In *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, Hong Kong, Apr.

T. Briscoe. 1996. Robust parsing. In R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue, editors, *Survey of the State of the Art of Human Language Technology*, chapter 3.7. Cambridge University Press, Cambridge, England.

CUData, 2004. *DARPA Communicator Travel Data*. University of Colorado at Boulder. http://communicator.colorado.edu/phoenix.

D.A. Dahl, M. Bates, M. Brown, K. Hunicke-Smith, D. Pallett, C. Pao, A. Rudnicky, and L. Shriberg. 1994. Expanding the scope of the ATIS task: the ATIS-3 corpus. In *ARPA Human Language Technology Workshop*, Princeton, NJ, Mar.

J. Dowding, R. Moore, F. Andry, and D. Moran. 1994. Interleaving syntax and semantics in an efficient bottom-up parser. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 110–116, Las Cruces, New Maxico, June.

N. Friedman, D. Geiger, and M. Goldszmidt. 1997. Bayesian network classifiers. *Machine Learning*, 29(2):131–163.

M.J. Gales and P.C. Woodland. 1996. Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 10:249–264, Oct.

J.L. Gauvain and C.-H. Lee. 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298.

V. Goel and W. Byrne. 1999. Task dependent loss functions in speech recognition: Application to named entity extraction. In *ESCA ETRW Workshop on Accessing Information from Spoken Audio*, pages 49–53, Cambridge, UK.

Yulan He and Steve Young. 2003a. A data-driven spoken language understanding system. In *IEEE Automatic Speech Recognition and Understanding Workshop*, St. Thomas, U.S. Virgin Islands, Dec.

Yulan He and Steve Young. 2003b. Hidden vector state model for hierarchical semantic parsing. In *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, Hong Kong, Apr.

HTK, 2004. *Hidden Markov Model Toolkit (HTK) 3.2*. Cambridge University Engineering Department. http://htk.eng.cam.ac.uk/.

D. Klakow. 1998. Log-linear interpolation of language models. In *Proc. of Intl. Conf. on Spoken Language Processing*, Sydney, Australia, Nov.

N. Kumar. 1997. *Investigation of Silicon Auditory Models and Generalization of Linear Discriminant analysis for Improved Speech Recognition*. Ph.D. thesis, Johns Hopkins University, Baltimore MD.

X. Luo, S. Roukos, and T. Ward. 1999. Unsupervised adaptation of statistical parsers based on Markov transform. In *IEEE Automatic Speech Recognition and Understanding Workshop*, Keystone, Colorado, Dec.

X. Luo. 2000. Parser adaptation via householder transform. In *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, June.

S. Martin, A. Kellner, and T. Portele. 2000. Interpolation of stochastic grammar and word bigram models in natural language understanding. In *Proc. of Intl. Conf. on Spoken Language Processing*, Beijing, China, Oct.

B. Roark and M. Bacchiani. 2003. Supervised and unsupervised PCFG adaptation to novel domains. In *Proceedings of the joint meeting of the North American Chapter of the Association for Computational Linguistics and the Human Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada, May.

S. Seneff. 1992. Robust parsing for spoken language systems. In *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, San Francisco.

A.P. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones. 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, DRA Speech Research Unit.

W. Ward and S. Issar. 1996. Recent improvements in the CMU spoken language understanding system. In *Proc. of the ARPA Human Language Technology Workshop*, pages 213–216. Morgan Kaufman Publishers, Inc.