Segmented and unsegmented dialogue-act annotation with statistical dialogue models*

Carlos D. Martínez Hinarejos, Ramón Granell, José Miguel Benedí
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera, s/n, 46022, Valencia
{cmartine,rgranell,jbenedi}@dsic.upv.es

Abstract

Dialogue systems are one of the most challenging applications of Natural Language Processing. In recent years, some statistical dialogue models have been proposed to cope with the dialogue problem. The evaluation of these models is usually performed by using them as annotation models. Many of the works on annotation use information such as the complete sequence of dialogue turns or the correct segmentation of the dialogue. This information is not usually available for dialogue systems. In this work, we propose a statistical model that uses only the information that is usually available and performs the segmentation and annotation at the same time. The results of this model reveal the great influence that the availability of a correct segmentation has in obtaining an accurate annotation of the dialogues.

1 Introduction

In the Natural Language Processing (NLP) field, one of the most challenging applications is dialogue systems (Kuppevelt and Smith, 2003). A dialogue system is usually defined as a computer system that can interact with a human being through dialogue in order to complete a specific task (e.g., ticket reservation, timetable consultation, bank operations,…) (Aust et al., 1995; Hardy et al., 2002). Most dialogue system have a characteristic behaviour with respect to dialogue management, which is known as dialogue strategy. It defines what the dialogue system must do at each point of the dialogue.

Most of these strategies are rule-based, i.e., the dialogue strategy is defined by rules that are usually defined by a human expert (Gorin et al., 1997; Hardy et al., 2003). This approach is usually difficult to adapt or extend to new domains where the dialogue structure could be completely different, and it requires the definition of new rules.

Similar to other NLP problems (like speech recognition and understanding, or statistical machine translation), an alternative data-based approach has been developed in the last decade (Stolcke et al., 2000; Young, 2000). This approach relies on statistical models that can be automatically estimated from annotated data, which in this case, are dialogues from the task.

Statistical modelling learns the appropriate parameters of the models from the annotated dialogues. As a simplification, it could be considered that each label is associated to a situation in the dialogue, and the models learn how to identify and react to the different situations by estimating the associations between the labels and the dialogue events (words, the speaker, previous turns, etc.). An appropriate annotation scheme should be defined to capture the elements that are really important for the dialogue, eliminating the information that is irrelevant to the dialogue process. Several annotation schemes have been proposed in the last few years (Core and Allen, 1997; Dybkjaer and Bernsen, 2000).

One of the most popular annotation schemes at the dialogue level is based on Dialogue Acts (DA). A DA is a label that defines the function of the annotated utterance with respect to the dialogue process. In other words, every turn in the dialogue

* Work partially supported by the Spanish project TIC2003-08681-C02-02 and by Spanish Ministry of Culture under FPI grants.

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 563–570,
Sydney, July 2006. ©2006 Association for Computational Linguistics

is supposed to be composed of one or more utterances. In this context, from the dialogue management viewpoint an utterance is a relevant subsequence . Several DA annotation schemes have been proposed in recent years (DAMSL (Core and Allen, 1997), VerbMobil (Alexandersson et al., 1998), Dihana (Alcácer et al., 2005)).

In all these studies, it is necessary to annotate a large amount of dialogues to estimate the parameters of the statistical models. Manual annotation is the usual solution, although is very time-consuming and there is a tendency for error (the annotation instructions are not usually easy to interpret and apply, and human annotators can commit errors) (Jurafsky et al., 1997).

Therefore, the possibility of applying statistical models to the annotation problem is really interesting. Moreover, it gives the possibility of evaluating the statistical models. The evaluation of the performance of dialogue strategies models is a difficult task. Although many proposals have been made (Walker et al., 1997; Fraser, 1997; Stolcke et al., 2000), there is no real agreement in the NLP community about the evaluation technique to apply.

Our main aim is the evaluation of strategy models, which provide the reaction of the system given a user input and a dialogue history. Using these models as annotation models gives us a possible evaluation: the correct recognition of the labels implies the correct recognition of the dialogue situation; consequently this information can help the system to react appropriately. Many recent works have attempted this approach (Stolcke et al., 2000; Webb et al., 2005).

However, many of these works are based on the hypothesis of the availability of the segmentation into utterances of the turns of the dialogue. This is an important drawback in order to evaluate these models as strategy models, where segmentation is usually not available. Other works rely on a decoupled scheme of segmentation and DA classification (Ang et al., 2005).

In this paper, we present a new statistical model that computes the segmentation and the annotation of the turns at the same time, using a statistical framework that is simpler than the models that have been proposed to solve both problems at the same time (Warnke et al., 1997). The results demonstrate that segmentation accuracy is really important in obtaining an accurate annotation of the dialogue, and consequently in obtaining quality strategy models. Therefore, more accurate segmentation models are needed to perform this process efficiently.

This paper is organised as follows: Section 2, presents the annotation models (for both the unsegmented and segmented versions); Section 3, describes the dialogue corpora used in the experiments; Section 4 establishes the experimental framework and presents a summary of the results; Section 5, presents our conclusions and future research directions.

2 Annotation models

The statistical annotation model that we used initially was inspired by the one presented in (Stolcke et al., 2000). Under a maximum likelihood framework, they developed a formulation that assigns DAs depending on the conversation evidence (transcribed words, recognised words from a speech recogniser, phonetic and prosodic features,...). Stolcke's model uses simple and popular statistical models: N-grams and Hidden Markov Models. The N-grams are used to model the probability of the DA sequence, while the HMM are used to model the evidence likelihood given the DA. The results presented in (Stolcke et al., 2000) are very promising.

However, the model makes some unrealistic assumptions when they are evaluated to be used as strategy models. One of them is that there is a complete dialogue available to perform the DA assignation. In a real dialogue system, the only available information is the information that is prior to the current user input. Although this alternative is proposed in (Stolcke et al., 2000), no experimental results are given.

Another unrealistic assumption corresponds to the availability of the segmentation of the turns into utterances. An utterance is defined as a dialogue-relevant subsequence of words in the current turn (Stolcke et al., 2000). It is clear that the only information given in a turn is the usual information: transcribed words (for text systems), recognised words, and phonetic/prosodic features (for speech systems). Therefore, it is necessary to develop a model to cope with both the segmentation and the assignation problem.

Let $U_1^d = U_1 U_2 \cdots U_d$ be the sequence of DA assigned until the current turn, corresponding to the first d segments of the current dialogue. Let

$W = w_1 w_2 \ldots w_l$ be the sequence of the words of the current turn, where subsequences $W_i^j = w_i w_{i+1} \ldots w_j$ can be defined ($1 \leq i \leq j \leq l$).

For the sequence of words W, a segmentation is defined as $s_1^r = s_0 s_1 \ldots s_r$, where $s_0 = 0$ and $W = W_{s_0+1}^{s_1} W_{s_1+1}^{s_2} \ldots W_{s_{r-1}+1}^{s_r}$. Therefore, the optimal sequence of DA for the current turn will be given by:

$$\hat{U} = \operatorname*{argmax}_{U} \Pr(U|W_1^l, U_1^d) =$$

$$\operatorname*{argmax}_{U_{d+1}^{d+r}} \sum_{(s_1^r, r)} \Pr(U_{d+1}^{d+r}|W_1^l, U_1^d)$$

After developing this formula and making several assumptions and simplifications, the final model, called *unsegmented model*, is:

$$\hat{U} = \operatorname*{argmax}_{U_{d+1}^{d+r}} \max_{(s_1^r, r)}$$

$$\prod_{k=d+1}^{d+r} \Pr(U_k|U_{k-n-1}^{k-1}) \Pr(W_{s_{k-(d+1)}+1}^{s_{k-d}}|U_k)$$

This model can be easily implemented using simple statistical models (N-grams and Hidden Markov Models). The decoding (segmentation and DA assignation) was implemented using the Viterbi algorithm. A Word Insertion Penalty (WIP) factor, similar to the one used in speech recognition, can be incorporated into the model to control the number of utterances and avoid excessive segmentation.

When the segmentation into utterances is provided, the model can be simplified into the *segmented model*, which is:

$$\hat{U} = \operatorname*{argmax}_{U_{d+1}^{d+r}}$$

$$\prod_{k=d+1}^{d+r} \Pr(U_k|U_{k-n-1}^{k-1}) \Pr(W_{s_{k-(d+1)}+1}^{s_{k-d}}|U_k)$$

All the presented models only take into account word transcriptions and dialogue acts, although they could be extended to deal with other features (like prosody, sintactical and semantic information, etc.).

3 Experimental data

Two corpora with very different features were used in the experiment with the models proposed in Section 2. The SwitchBoard corpus is composed of human-human, non task-oriented dialogues with a large vocabulary. The Dihana corpus is composed of human-computer, task-oriented dialogues with a small vocabulary.

Although two corpora are not enough to let us draw general conclusions, they give us more reliable results than using only one corpus. Moreover, the very different nature of both corpora makes our conclusions more independent from the corpus type, the annotation scheme, the vocabulary size, etc.

3.1 The SwitchBoard corpus

The first corpus used in the experiments was the well-known SwitchBoard corpus (Godfrey et al., 1992). The SwitchBoard database consists of human-human conversations by telephone with no directed tasks. Both speakers discuss about general interest topics, but without a clear task to accomplish.

The corpus is formed by 1,155 conversations, which comprise 126,754 different turns of spontaneous and sometimes overlapped speech, using a vocabulary of 21,797 different words. The corpus was segmented into utterances, each of which was annotated with a DA following the simplified DAMSL annotation scheme (Jurafsky et al., 1997). The set of labels of the simplified DAMSL scheme is composed of 42 different labels, which define categories such as statement, backchannel, opinion, etc. An example of annotation is presented in Figure 1.

3.2 The Dihana corpus

The second corpus used was a task-oriented corpus called Dihana (Benedí et al., 2004). It is composed of computer-to-human dialogues, and the main aim of the task is to answer telephone queries about train timetables, fares, and services for long-distance trains in Spanish. A total of 900 dialogues were acquired by using the Wizard of Oz technique and semicontrolled scenarios. Therefore, the voluntary caller was always free to express him/herself (there were no syntactic or vocabulary restrictions); however, in some dialogues, s/he had to achieve some goals using a set of restrictions that had been given previously (e.g. departure/arrival times, origin/destination, travelling on a train with some services, etc.).

These 900 dialogues comprise 6,280 user turns and 9,133 system turns. Obviously, as a task-

Utterance	Label
YEAH, TO GET REFERENCES AND THAT, SO, BUT, UH, I DON'T FEEL COMFORTABLE ABOUT LEAVING MY KIDS IN A BIG DAY CARE CENTER, SIMPLY BECAUSE THERE'S SO MANY KIDS AND SO MANY <SNIFFING> <THROAT_CLEARING>	
Yeah,	aa
to get references and that,	sd
so, but, uh,	%
I don't feel comfortable about leaving my kids in a big day care center, simply because there's so many kids and so many <sniffing> <throat_clearing>	sd
I THINK SHE HAS PROBLEMS WITH THAT, TOO.	
I think she has problems with that, too.	sd

Figure 1: An example of annotated turns in the SwitchBoard corpus.

oriented and medium size corpus, the total number of different words in the vocabulary, 812, is not as large as the Switchboard database.

The turns were segmented into utterances. It was possible for more than one utterance (with their respective labels) to appear in a turn (on average, there were 1.5 utterances per user/system turn). A three-level annotation scheme of the utterances was defined (Alcácer et al., 2005). These labels represent the general purpose of the utterance (first level), as well as more specific semantic information (second and third level): the second level represents the data focus in the utterance and the third level represents the specific data present in the utterance. An example of three-level annotated user turns is given in Figure 2. The corpus was annotated by means of a semiautomatic procedure, and all the dialogues were manually corrected by human experts using a very specific set of defined rules.

After this process, there were 248 different labels (153 for user turns, 95 for system turns) using the three-level scheme. When the detail level was reduced to the first and second levels, there were 72 labels (45 for user turns, 27 for system turns). When the detail level was limited to the first level, there were only 16 labels (7 for user turns, 9 for system turns). The differences in the number of labels and in the number of examples for each label with the SwitchBoard corpus are significant.

4 Experiments and results

The SwitchBoard database was processed to remove certain particularities. The main adaptations performed were:

- The interrupted utterances (which were labelled with '+') were joined to the correct previous utterance, thereby avoiding interruptions (i.e., all the words of the interrupted utterance were annotated with the same DA).

Table 1: SwitchBoard database statistics (mean for the ten cross-validation partitions)

	Training	Test
Dialogues	1,136	19
Turns	113,370	1,885
Utterances	201,474	3,718
Running words	1,837,222	33,162
Vocabulary	21,248	2,579

- All the words were transcribed in lowercase.
- Puntuaction marks were separated from words.

The experiments were performed using a cross-validation approach to avoid the statistical bias that can be introduced by the election of fixed training and test partitions. This cross-validation approach has also been adopted in other recent works on this corpus (Webb et al., 2005). In our case, we performed 10 different experiments. In each experiment, the training partition was composed of 1,136 dialogues, and the test partition was composed of 19 dialogues. This proportion was adopted so that our results could be compared with the results in (Stolcke et al., 2000), where similar training and test sizes were used. The mean figures for the training and test partitions are shown in Table 1.

With respect to the Dihana database, the preprocessing included the following points:

- A categorisation process was performed for categories such as town names, the time, dates, train types, etc.
- All the words were transcribed in lowercase.
- Puntuaction marks were separated from words.
- All the words were preceded by the speaker identification (U for user, M for system).

Utterance	1st level	2nd level	3rd level
YES, TIMES AND FARES.			
Yes, times and fares	Acceptance Question	Dep_Hour Dep_Hour,Fare	Nil Nil
YES, I WANT TIMES AND FARES OF TRAINS THAT ARRIVE BEFORE SEVEN.			
Yes, I want times and fares of trains that arrive before seven.	Question	Dep_Hour,Fare	Arr_Hour
ON THURSDAY IN THE AFTERNOON.			
On thursday in the afternoon	Answer Answer	Day Time	Day Time

Figure 2: An example of annotated turns in the Dihana corpus. Original turns were in Spanish.

Table 2: Dihana database statistics (mean for the five cross-validation partitions)

	Training	Test
Dialogues	720	180
Turns	12,330	3,083
User turns	5,024	1,256
System turns	7,206	1,827
Utterances	18,837	4,171
User utterances	7,773	1,406
System utterances	11,064	2,765
Running words	162,613	40,765
User running words	42,806	10,815
System running words	119,807	29,950
Vocabulary	832	485
User vocabulary	762	417
System vocabulary	208	174

Table 3: SwitchBoard results for the segmented model

N-gram	Utt. accuracy	Turn accuracy
2-gram	68.19%	59.33%
3-gram	68.50%	59.75%
4-gram	67.90%	59.14%

A cross-validation approach was adopted in Dihana as well. In this case, only 5 different partitions were used. Each of them had 720 dialogues for training and 180 for testing. The statistics on the Dihana corpus are presented in Table 2.

For both corpora, different N-gram models, with $\mathrm{N} = 2, 3, 4$, and HMM of one state were trained from the training database. In the case of the SwitchBoard database, all the turns in the test set were used to compute the labelling accuracy. However, for the Dihana database, only the user turns were taken into account (because system turns follow a regular, template-based scheme, which presents artificially high labelling accuracies). Furthermore, in order to use a really significant set of labels in the Dihana corpus, we performed the experiments using only two-level labels instead of the complete three-level labels. This restriction allowed us to be more independent from the understanding issues, which are strongly related to the third level. It also allowed us to concentrate on the dialogue issues, which relate more to the first and second levels.

The results in the case of the segmented approach described in Section 2 for SwitchBoard are presented in Table 3. Two different definitions of accuracy were used to assess the results:

- Utterance accuracy: computes the proportion of well-labelled utterances.
- Turn accuracy: computes the proportion of totally well-labelled turns (i.e.: if the labelling has the same labels in the same order as in the reference, it is taken as a well-labelled turn).

As expected, the utterance accuracy results are a bit worse than those presented in (Stolcke et al., 2000). This may be due to the use of only the past history and possibly to the cross-validation approach used in the experiments. The turn accuracy was calculated to compare the segmented and the unsegmented models. This was necessary because the utterance accuracy does not make sense for the unsegmented model.

The results for the unsegmented approach for SwitchBoard are presented in Table 4. In this case, three different definitions of accuracy were used to assess the results:

- Accuracy at DA level: the edit distance between the reference and the labelling of the turn was computed; then, the number of correct substitutions (c), wrong substitutions (s), deletions (d) and insertions (i) was com-

Table 4: SwitchBoard results for the unsegmented model (WIP=50)

N-gram	DA acc.	Turn acc.	Segm. acc.
2-gram	38.19%	39.47%	38.92%
3-gram	38.58%	39.61%	39.06%
4-gram	38.49%	39.52%	38.96%

puted, and the accuracy was calculated as $100 \cdot \frac{c}{(c+s+i+d)}$.

- Accuracy at turn level: this provides the proportion of well-labelled turns, without taking into account the segmentation (i.e., if the labelling has the same labels in the same order as in the reference, it is taken as a well-labelled turn).
- Accuracy at segmentation level: this provides the proportion of well-labelled and segmented turns (i.e., the labels are the same as in the reference and they affect the same utterances).

The WIP parameter used in Table 4 was 50, which is the one that offered the best results. The segmentation accuracy in Table 4 must be compared with the turn accuracy in Table 3. As Table 4 shows, the accuracy of the labelling decreased dramatically. This reveals the strong influence of the availability of the real segmentation of the turns.

To confirm this hypothesis, similar experiments were performed with the Dihana database. Table 5 presents the results with the segmented corpus, and Table 6 presents the results with the unsegmented corpus (with WIP=50, which gave the best results). In this case, only user turns were taken into account to compute the accuracy, although the model was applied to all the turns (both user and system turns). For the Dihana corpus, the degradation of the results of the unsegmented approach with respect to the segmented approach was not as high as in the SwitchBoard corpus, due to the smaller vocabulary and complexity of the dialogues.

These results led us to the same conclusion, even for such a different corpus (much more labels, task-oriented, etc.). In any case, these accuracy figures must be taken as a lower bound on the model performance because sometimes an incorrect recognition of segment boundaries or dialogue acts does not cause an inappropriate reaction of the dialogue strategy.

Table 5: Dihana results for the segmented model (only two-level labelling for user turns)

N-gram	Utt. accuracy	Turn accuracy
2-gram	75.70%	74.46%
3-gram	76.28%	74.93%
4-gram	76.39%	75.10%

Table 6: Dihana results for the unsegmented model (WIP=50, only two-level labelling for user turns)

N-gram	DA acc.	Turn acc.	Segm. acc.
2-gram	60.36%	62.86%	58.15%
3-gram	60.05%	62.49%	57.87%
4-gram	59.81%	62.44%	57.88%

An illustrative example of annotation errors in the SwitchBoard database, is presented in Figure 3 for the same turns as in Figure 1. An error analysis of the segmented model was performed. The results reveals that, in the case of most of the errors were produced by the confusion of the ’sv’ and ’sd’ classes (about 50% of the times ’sv’ was badly labelled, the wrong label was ’sd’) The second turn in Figure 3 is an example of this type of error. The confusions between the ’aa’ and ’b’ classes were also significant (about 27% of the times ’aa’ was badly labelled, the wrong label was ’b’). This was reasonable due to the similar definitions of these classes (which makes the annotation difficult, even for human experts). These errors were similar for all the N-grams used. In the case of the unsegmented model, most of the errors were produced by deletions of the ’sd’ and ’sv’ classes, as in the first turn in Figure 3 (about 50% of the errors). This can be explained by the presence of very short and very long utterances in both classes (i.e., utterances for ’sd’ and ’sv’ did not present a regular length).

Some examples of errors in the Dihana corpus are shown in Figure 4 (in this case, for the same turns as those presented in Figure 2). In the segmented model, most of the errors were substitutions between labels with the same first level (especially questions and answers) where the second level was difficult to recognise. The first and third turn in Figure 4 are examples of this type of error. This was because sometimes the expressions only differed with each other by one word, or

Utt	Label	
1	%	Yeah, to get references and that, so, but, uh, I don't
2	sd	feel comfortable about leaving my kids in a big day care center, simply because there's so many kids and so many <sniffing> <throat_clearing>
Utt	Label	
1	sv	I think she has problems with that, too.

Figure 3: An example of errors produced by the model in the SwitchBoard corpus

the previous segment influence (i.e., the language model weight) was not enough to get the appropriate label. This was true for all the N-grams tested. In the case of the unsegmented model, most of the errors were caused by similar misrecognitions in the second level (which are more frequent due to the absence of utterance boundaries); however, deletion and insertion errors were also significant. The deletion errors corresponded to acceptance utterances, which were too short (most of them were "Yes"). The insertion errors corresponded to "Yes" words that were placed after a new-consult system utterance, which is the case of the second turn presented in Figure 4. These words should not have been labelled as a separate utterance. In both cases, these errors were very dependant on the WIP factor, and we had to get an adequate WIP value which did not increase the insertions and did not cause too many deletions.

5 Conclusions and future work

In this work, we proposed a method for simultaneous segmentation and annotation of dialogue utterances. In contrast to previous models for this task, our model does not assume manual utterance segmentation. Instead of treating utterance segmentation as a separate task, the proposed method selects utterance boundaries to optimize the accuracy of the generated labels. We performed experiments to determine the effect of the availability of the correct segmentation of dialogue turns in utterances in the statistical DA labelling framework. Our results reveal that, as shown in previous work (Warnke et al., 1999), having the correct segmentation is very important in obtaining accurate results in the labelling task. This conclusion is supported by the results obtained in very different dialogue corpora: different amounts of training and test data, different natures (general and task-oriented), different sets of labels, etc.

Future work on this task will be carried out in several directions. As segmentation appears to be an important step in these tasks, it would be interesting to obtain an automatic and accurate segmentation model that can be easily integrated in our statistical model. The application of our statistical models to other tasks (like VerbMobil (Alexandersson et al., 1998)) would allow us to confirm our conclusions and compare results with other works.

The error analysis we performed shows the need for incorporating new and more reliable information resources to the presented model. Therefore, the use of alternative models in both corpora, such as the N-gram-based model presented in (Webb et al., 2005) or an evolution of the presented statistical model with other information sources would be useful. The combination of these two models might be a good way to improve results.

Finally, it must be pointed out that the main task of the dialogue models is to allow the most correct reaction of a dialogue system given the user input. Therefore, the correct evaluation technique must be based on the system behaviour as well as on the accurate assignation of DA to the user input. Therefore, future evaluation results should take this fact into account.

Acknowledgements

The authors wish to thank Nick Webb, Mark Hepple and Yorick Wilks for their comments and suggestions and for providing the preprocessed SwitchBoard corpus. We also want to thank the anonymous reviewers for their criticism and suggestions.

References

N. Alcácer, J. M. Benedí, F. Blat, R. Granell, C. D. Martínez, and F. Torres. 2005. Acquisition and labelling of a spontaneous speech dialogue corpus. In *Proceedings of SPECOM*, pages 583–586, Patras, Greece.

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elis-

Utterance	1st level	2nd level
Yes, times	Acceptance	Dep_Hour,Fare
and fares	Question	Dep_Hour,Fare
Yes, I want	Acceptance	Dep_Hour,Fare
times and fares of trains that arrive before seven.	Question	Dep_Hour,Fare
On thursday in the afternoon	Answer	Time

Figure 4: An example of errors produced by the model in the Dihana corpus

abeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1998. Dialogue acts in VERBMOBIL-2 (second edition). Technical Report 226, DFKI GmbH, Saarbrücken, Germany, July.

J. Ang, Y. Liu, and E. Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the International Conference of Acoustics, Speech, and Signal Processings*, volume 1, pages 1061–1064, Philadelphia.

H. Aust, M. Oerder, F. Seide, and V. Steinbiss. 1995. The philips automatic train timetable information system. *Speech Communication*, 17:249–263.

J. M. Benedí, A. Varona, and E. Lleida. 2004. Dihana: Dialogue system for information access using spontaneous speech in several environments tic2002-04103-c03. In *Reports for Jornadas de Seguimiento - Programa Nacional de Tecnologías Informáticas*, Málaga, Spain.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the damsl annotation scheme. In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA, November.

Layla Dybkjaer and Niels Ole Bernsen. 2000. The mate workbench.

N. Fraser, 1997. *Assessment of interactive systems*, pages 564–614. Mouton de Gruyter.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proc. ICASSP-92*, pages 517–520.

A. Gorin, G. Riccardi, and J. Wright. 1997. How may i help you? *Speech Communication*, 23:113–127.

Hilda Hardy, Kirk Baker, Laurence Devillers, Lori Lamel, Sophie Rosset, Tomek Strzalkowski, Cristian Ursu, and Nick Webb. 2002. Multi-layer dialogue annotation for automated multilingual customer service. In *Proceedings of the ISLE Workshop on Dialogue Tagging for Multi-Modal Human Computer Interaction*, Edinburgh, Scotland, December.

Hilda Hardy, Tomek Strzalkowski, and Min Wu. 2003. Dialogue management for an automated multilingual call center. In *Proceedings of HLT-NAACL 2003 Workshop: Research Directions in Dialogue Processing*, pages 10–12, Edmonton, Canada, June.

D. Jurafsky, E. Shriberg, and D. Biasca. 1997. Switchboard swbd-damsl shallow- discourse-function annotation coders manual - draft 13. Technical Report 97-01, University of Colorado Institute of Cognitive Science.

J. Van Kuppevelt and R. W. Smith. 2003. *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*. Springer.

A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. 2000. Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):1–34.

Marilyn A. Walker, Diane Litman J., Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Somerset, New Jersey. Association for Computational Linguistics.

V. Warnke, R. Kompe, H. Niemann, and E. Nöth. 1997. Integrated Dialog Act Segmentation and Classification using Prosodic Features and Language Models. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 207–210, Rhodes.

V. Warnke, S. Harbeck, E. Nöth, H. Niemann, and M. Levit. 1999. Discriminative Estimation of Interpolation Parameters for Language Model Classifiers. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 525–528, Phoenix, AZ, March.

N. Webb, M. Hepple, and Y. Wilks. 2005. Dialogue act classification using intra-utterance features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding*, Pittsburgh.

S. Young. 2000. Probabilistic methods in spoken dialogue systems. *Philosophical Trans Royal Society (Series A)*, 358(1769):1389–1402.