

Evaluation of HMM-based models for the annotation of unsegmented dialogue turns

Carlos-D. Martínez-Hinarejos, Vicent Tamarit, José-Miguel Benedí

Institut Tecnològic d'Informàtica, Universitat Politècnica de València
Camí de Vera, s/n. 46022 València, Spain
{cmartine,vtamarit,jbenedi}@dsic.upv.es

Abstract

Corpus-based dialogue systems rely on statistical models, whose parameters are inferred from annotated dialogues. The dialogues are usually annotated using Dialogue Acts (DA), and the manual annotation is difficult and time-consuming. Therefore, several semi-automatic annotation processes have been proposed to speed-up the process. The standard annotation model is based on Hidden Markov Models (HMM). In this work, we explore the impact of different types of HMM on annotation accuracy using these models on two dialogue corpora of dissimilar features. The results show that some types of models improve standard HMM in a human-computer task-oriented dialogue corpus, but their impact is lower in a human-human non-task-oriented dialogue corpus.

1. Introduction

A dialogue system is a natural language application where a user asks a computer system for some information and some interaction using dialogue is needed to get the required information (Dybkjær and Minker, 2008). The system reacts to user interactions using the so-called dialogue strategy. A dialogue strategy can be defined by rules (rule-based approach) (Gorin et al., 1997; Hardy et al., 2003) or by statistical models obtained from acquired dialogues (data-based approach) (Young, 2000). The parameters of data-based dialogue statistical models can be inferred from annotated dialogues. Dialogues are usually annotated in the form of Dialogue Acts (DA) (Bunt, 1994).

A DA is a label that expresses the intention and function of the corresponding dialogue segment. From the viewpoint of the dialogue, a segment is the minimal informational unit that is assigned to a single DA. Each dialogue turn may have one or more segments. The annotation of a complete dialogue corpus is required to infer the parameters of the statistical models. Consequently, the annotation task is an important step in the development of dialogue systems.

The manual annotation of dialogues is hard, time-consuming, and error-prone. Therefore, the use of semi-automatic annotation tools is really interesting to speed-up this process, as these tools can provide a draft annotation that can be reviewed by a human annotator in less time than annotating from scratch. Some probabilistic models have been proposed for this annotation (Stolcke et al., 2000; Webb and Wilks, 2005). Many of them assume the previous segmentation of the turns into segments, which is not usual in the transcribed dialogue corpora that are available. Recently, some models have been proposed to directly annotate unsegmented dialogues (Martínez-Hinarejos et al., 2008). Other authors propose a statistical segmentation previous to the annotation process (Ang et al., 2005).

The most widely used probabilistic models in this field are Hidden Markov Models (HMM) combined with N-grams of DA (Stolcke et al., 2000). In the annotation task on unsegmented dialogue turns, it is important to obtain the correct segmentation and the correct DA label for each segment. Previous works (Martínez-Hinarejos et al., 2009)

showed a poor segmentation accuracy with standard one-state-with-loop topology HMM. In this work, we evaluate the segmentation and annotation accuracy of other types of HMM (left-to-right with loops with more than one state). The evaluation is performed on two dialogue corpora with very different features in order to validate the performance of the models in different situations.

This paper is organised as follows: In Section 2. we introduce the model. In Section 3. we present the corpora. In Section 4., we report the experiments and results. In Section 5., we present the conclusions and future work lines.

2. The HMM-based Annotation Model

The problem of obtaining the optimal sequence of DA \mathcal{U} corresponding to the sequence of words of the dialogue \mathcal{W} can be stated as:

$$\hat{\mathcal{U}} = \underset{\mathcal{U}}{\operatorname{argmax}} \Pr(\mathcal{U}|\mathcal{W}) \quad (1)$$

We can express the complete sequences of words and DA in terms of the different turns in the dialogue: given a dialogue with T turns, its associated word sequence and DA sequence are $\mathcal{W} = W_1^T = W_1 W_2 \cdots W_T$ and $\mathcal{U} = U_1^T = U_1 U_2 \cdots U_T$, respectively. Thus, we can express (1) as:

$$\hat{\mathcal{U}} = \underset{\mathcal{U}}{\operatorname{argmax}} \Pr(\mathcal{U}|\mathcal{W}) = \underset{U_1^T}{\operatorname{argmax}} \Pr(U_1^T|W_1^T) \quad (2)$$

Using the Bayes rule, (2) can be formulated as:

$$\begin{aligned} \underset{U_1^T}{\operatorname{argmax}} \Pr(U_1^T|W_1^T) &= \underset{U_1^T}{\operatorname{argmax}} \Pr(U_1^T) \Pr(W_1^T|U_1^T) = \\ &\approx \underset{U_1^T}{\operatorname{argmax}} \prod_{t=1}^T \Pr(U_t|U_1^{t-1}) \Pr(W_t|W_1^{t-1}, U_1^T) \\ &\approx \underset{U_1^T}{\operatorname{argmax}} \prod_{t=1}^T \Pr(U_t|U_1^{t-1}) \Pr(W_t|U_1^t) \end{aligned} \quad (3)$$

The approximation in (3) is based on two reasonable assumptions:

- Word sequences of turn t are only affected by the DA sequences of previous turns, and not by future turns.
- As W_1^T is the sequence of given events, dependencies between word sequences can be ignored.

Previous works (Stolcke et al., 2000) have proposed similar approaches to DA annotation, but they assume the availability of the segmentation of the turn (Stolcke et al., 2000; Webb and Wilks, 2005). In our case, we develop the formulation to use the model in the unsegmented case as follows:

- $W_t = w_1^l = w_1 w_2 \dots w_l$ is described in terms of all the possible segmentations as $W_t = w_{s_0+1}^{s_1} w_{s_1+1}^{s_2} \dots w_{s_{r-1}+1}^{s_r}$, where r is the number of segments and s_k is the index of the k -th segment, with $s_0 = 0$ and $s_r = l$.
- $U_t = u_1^r$ is the DA sequence of turn t .
- $U_1^{t-1} = U_1 U_2 \dots U_{t-1}$ are the DA sequences previous to turn t .

Therefore, using $W = W_t$ and $U = U_t$ to simplify notation, the terms in the product in (3) can be expressed as:

$$\Pr(U|U_1^{t-1}) \Pr(W|U_1^t) = \sum_{r,s_1^r} \prod_{k=1}^r \Pr(u_k|u_1^{k-1}, U_1^{t-1}) \Pr(w_{s_{k-1}+1}^{s_k}|u_1^k, U_1^{t-1}) \quad (4)$$

Notice that in this reformulation, W and U (capital letters) represent sequences of words and DA, respectively, while w and u (lowercase letters) represent single words and DA, respectively.

This model can be simplified with some assumptions:

- The current DA depends only on the previous $n - 1$ DAs: $\Pr(u_k|u_1^{k-1}, U_1^{t-1}) \approx \Pr(u_k|u_{k-n+1}^{k-1})$.
- The sequence of words of the current segment depends only on the current DA: $\Pr(w_{s_{k-1}+1}^{s_k}|u_1^k, U_1^{t-1}) \approx \Pr(w_{s_{k-1}+1}^{s_k}|u_k)$.

From this model, we can formulate the search problem. This search looks for the maximum probability segmentation and DA sequence, and it is solved using the Viterbi process on the whole dialogue. Thus, the implementation of the search process results in the following final model:

$$\hat{U} = \operatorname{argmax}_{\mathcal{U}} \max_{R,S} \prod_{t=1}^T \prod_{k=1}^r \Pr(u_k|u_{k-n+1}^{k-1}) \Pr(w_{s_{k-1}+1}^{s_k}|u_k) \quad (5)$$

In this formula, $R = \{r_1, r_2, \dots, r_T\}$ represents the set of number of segments for each turn, and $S = \{s_1^{r_1}, s_1^{r_2}, \dots, s_1^{r_T}\}$ is the set of segmentations for each turn that maximises the product. Note that to simplify notation, the terms in the product are defined in terms of $r = r_t$ and $s_1^r = s_1^{r_t}$.

The models employed in the implementation of the search problem given by (5) are the following:

- $\Pr(u_k|u_{k-n+1}^{k-1})$ is represented by a statistical model of DA sequences (DA language model), generally an n-gram model.
- $\Pr(w_{s_{k-1}+1}^{s_k}|u_k)$ is modelled by a HMM.

In the model represented by (5), the contribution of the DA language model can be balanced by using a weight factor. This model was presented and tested in (Martínez-Hinarejos et al., 2009), where the topology of the HMM was one state with loop. This topology was imposed by the fact that the shortest possible segment has only one word, and, consequently, cannot be accepted by a HMM topology of more than one state. However, this topology had the side-effect of the loss of the word order, which is fundamental in order to obtain appropriate segmentations. This is caused by the fact that some words (especially punctuation marks and function words) are common to all segments (independently of the DA associated to that segment), and they are common segment delimiters as well. Therefore, words of this kind can be assigned to the next or previous segment without altering the decoding probability since the word probability is similar in all the HMM and they have a single state. However, this causes a dramatic decrease in the segmentation accuracy.

Therefore, we performed the same experiments with different HMM in order to study their impact on the annotation and segmentation accuracy. We used left-to-right-with-loops HMM, with more than one state, all of which were final states.

3. Corpora

We performed experiments on two very different dialogue corpora in order to obtain an appropriate view of the general impact of the new HMM on the annotation task. Both corpora are spontaneous-speech telephone-based corpora, with a moderate number of DA labels, but they differ in language, total size, task-orientation, vocabulary size, and the nature of the speakers (human-computer vs. human-human). A summary of the different features of the corpora is presented in Table 1.

3.1. SwitchBoard Corpus

The SwitchBoard corpus (Godfrey et al., 1992) is a well-known corpus of human-to-human telephone conversations in English. The conversations are about general topics, with no clear task to accomplish. This corpus recorded spontaneous speech, with frequent interruptions between the speakers, hesitations, non-linguistic sounds (laughter, coughing) and background noises. The transcription of the SwitchBoard took into account all these phenomena, including a special coding for each one. We must point out that SwitchBoard is a sort of standard corpus for evaluating automatic dialogue annotation tools, but it does not represent the usual aim of a dialogue system (task-oriented and human-computer dialogues).

The corpus consists of 1,155 conversations, with approximately 115,000 different turns. The vocabulary size is about 42,000 words. The dialogue annotation was performed using the SWBD-DAMSL scheme (Jurafsky et al.,

Table 1: Summary of the different features of the SwitchBoard and the Dihana corpus.

Corpus	SwitchBoard	DIHANA
Language	English	Spanish
Nature	Human-human	Human-computer
Task-oriented	No	Yes
Number of dialogues	1,155	900
Number of turns	115,000	6,280 U + 9,133 S
Annotation scheme	SWBD-DAMSL	IF-Dihana (3 and 2 levels)
Number of DA labels	42	153 U + 95 S / 45 U + 27 S

1997), which contains 42 different labels. The SWBD-DAMSL set is a simplified version of the general DAMSL annotation scheme (Core and Allen, 1997). SWBD-DAMSL includes several labels for different communicative categories at the dialogue level, such as statement, question, backchannels, etc., and more specific subcategories, such as statement-opinion, statement-non-opinion, yes-no-question, open-question, etc. In the annotation process, each turn was divided into several segments and each segment was assigned a DA label by a human annotator, according a predefined set of rules. The annotation was performed by 8 different human experts, with a Kappa value of 0.80 (Stolcke et al., 2000).

To simplify the experimental framework, we semi-automatically preprocessed the SwitchBoard corpus to remove certain phenomena: interruptions and overlaps were erased (by joining the interrupted turns), all the words were transcribed to lowercase, and punctuation marks were separated from the words. This preprocessed version of the SwitchBoard corpus is available at (Martínez-Hinarejos, 2010).

3.2. Dihana Corpus

The Dihana corpus is a spontaneous-speech Spanish-language, task-oriented corpus about the consultation of timetables and fares for trains (Benedí et al., 2004). The dialogue corpus was acquired using the Wizard of Oz (WoZ) (Fraser and Gilbert, 1991) set-up. A total number of 900 dialogues were acquired for the project, with a total of more than 15,000 turns (6,280 for user turns and 9,133 for system turns). A total number of 225 voluntary speakers participated in the acquisition process. Dihana is a more representative corpus of usual dialogue systems since it represents a set of task-oriented dialogues in a human-computer framework.

These dialogues were manually transcribed and annotated at the dialogue level by a set of human annotators, and it was consistently reviewed by a single annotator. The DA label set was defined from the Interchange Format (IF) proposal (Fukada et al., 1998), which defines the labels with three different levels: speech act, concept, and argument. The Dihana annotation set definition uses the first level to represent the general purpose of the segment and the second and third levels to represent more precise and task-oriented information (such as data repository or specific data contained in the segment).

The set of DA labels was composed of 248 different 3-

level labels (153 for user segments and 95 for system segments) (Alcácer et al., 2005). The high specificity of the third level leads to an alternative annotation using only the first level and the second level. In this case, there were 72 different 2-level labels (45 for user segments and 27 for system segments). Both versions of the annotated corpus were used in the experiments.

The final vocabulary was composed of 980 words. Before applying the annotation technique, the following preprocessing steps were performed: categorisation (e.g., town names, hours, dates, ...); identification of the speaker for each word (which is trivially obtained from the transcription of the dialogues since each turn is identified as being from the user or the wizard); lowercase transcription and separation of punctuation marks. The categorisation was automatically performed by using regular-expression-based tools that were manually coded and tested by a human expert.

4. Experiments and Results

The experiments evaluated the accuracy of the annotation and segmentation given by the models. To obtain more reliable results, we performed the experiments with a cross-validation approach. In our case, for the SwitchBoard corpus we used 11 partitions (105 dialogues each partition) for the Dihana corpus, we used 5 partitions (180 dialogues for each partition). The weight factor for the DA language model was optimised with respect to the global cross-validation experiment.

We used the following evaluation metrics, comparing the results turn by turn:

- DAER: the average edit distance between the reference DA sequences and the annotation results.
- SegER: the average edit distance between the reference segmentation (position of the last word of the segment) and the segmentation given by the model.
- SegDAER: the average edit distance between the combination of the DA and the segmentation in the reference and the annotation results.

Although other evaluation metrics can be adopted (such as those proposed in (Ang et al., 2005)), our proposed metrics are suitable for the evaluation of the annotation accuracy in the framework of dialogue annotation. This is because both the DA label and the position could be corrected in the

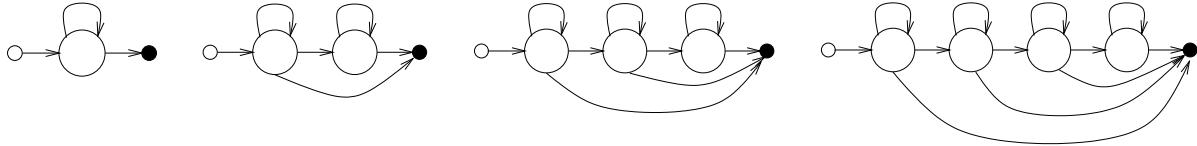


Figure 1: Different HMM topologies used in the decoding experiments. The small white circles denote artificial initial, and the small black circles denote artificial final states.

Table 2: SwitchBoard results for the different HMM topologies. The best results are in boldface.

DA N-gram	3-gram			4-gram			
	Num. states	DAER	SegER	SegDAER	DAER	SegER	SegDAER
1 (Baseline)	55.5	41.8	61.8	55.8	41.7	62.0	
2	57.1	37.3	65.8	58.3	37.6	67.0	
3	51.7	35.6	62.2	52.4	35.9	63.1	
4	54.2	36.0	64.3	55.1	36.3	65.4	

draft annotation. Consequently, the measures that take into account the difference in the labels (DAER), the difference in the positions (SegER), and the difference in both terms (SegDAER) are all good enough for the evaluation.

The topologies of the HMM are left-to-right with loops, and every state has a probability greater than zero of being a final state (which is given by the training process or by using a smoothing factor of 10^{-5} when the training process produces zero probability). The number of states went from one (baseline) to four, as can be seen in Figure 1. Tests with a larger number of states gave very inaccurate results due to data sparsity.

The results for the SwitchBoard corpus with 3-grams and 4-grams as the DA language models and the proposed HMM topologies are presented in Table 2. In this case, as the table shows, there is no improvement in SegDAER with respect to the baseline one-state topology. Although DAER and SegER improve with the three-state topology, SegDAER does not improve since the correct segmentation does not imply a correct label in the segments and vice-versa. In this case, the sparsity of the vocabulary of the corpus makes the new topologies unable to improve the baseline results because output distributions cannot be accurately estimated from such sparse data. Consequently, as the number of states grows, the number of output parameters grows in the same proportion, dramatically increasing the sparseness problem. Thus, the new topologies do not improve the results in dialogues with a large vocabulary, even though the number of DA labels is quite low.

The results for the Dihana corpus under the same conditions for the 3-level labels are presented in Table 3. These results show that the use of the new topologies have a beneficial effect on the segmentation of the turns, since the SegER and SegDAER results dramatically decrease. This is caused by the minimal word order that is modelled by a more-than-one-state topology: the one-state-topology may assign the symbols that are usually at the end of the segment to the beginning of another segment, whereas the other topologies with more than one state prevent this incorrect assignment since the initial states do not emit these symbols. Ta-

ble 4 presents the results for the 2-level labels. These results show a similar behaviour of the models, but with even more dramatic decreases in segmentation errors.

The best results for the Dihana corpus were obtained with a two-state topology in both 2-level and 3-level labels. A possible reason for this is that this topology has a lower number of parameters to estimate (only two output distributions by HMM) but it allows a minimal word order of the segment to be maintained. All HMM with a higher number of states showed a performance degradation since they did not contribute to a better word order and they also present a higher number of parameters to estimate.

A comparison of these results with the results given by another technique (the NGT technique (Martínez-Hinarejos et al., 2009)) show that the new topologies present similar results to NGT in the Dihana corpus (the best results with NGT in SegDAER were 17.9 and 8.3 for 3-level and 2-level labels, respectively) because of its structured nature and clear transcription. However, for the SwitchBoard corpus, the new topologies provide poorer results than the NGT technique (the best NGT result in this case was a SegDAER of 50.4).

Therefore, we can conclude that the new HMM models are appropriate for task-oriented and well-transcribed corpora (using the appropriate HMM topology), but they are not well fitted when applied to more unstructured corpora that have a higher level of vocabulary sparseness due to their non-task-oriented nature. In any case, as the development of dialogue systems is directed towards task-oriented systems, the proposed HMM-based model with the alternative HMM topologies seems to be a good alternative for the annotation of dialogue corpora for future systems, allowing the construction of dialogue systems to be performed faster.

5. Conclusions and Future Work

In this paper, we tested the effect of using different HMM topologies in a statistical model for dialogue annotation. The differences between the topologies are related to the number of states of the HMM since all the topologies have in common that they are left-to-right with loops and all

Table 3: Dihana 3-level results for the different HMM topologies. The best results are in boldface.

DA N-gram	3-gram			4-gram		
	Num. states	DAER	SegER	SegDAER	DAER	SegER
1 (Baseline)	15.6	26.2	34.5	15.9	26.2	34.7
2	16.2	7.9	17.8	16.7	8.0	18.4
3	17.4	9.0	19.2	18.0	9.4	20.0
4	20.4	8.4	22.0	20.7	8.6	22.4

Table 4: Dihana 2-level results for the different HMM topologies. The best results are in boldface.

DA N-gram	3-gram			4-gram		
	Num. states	DAER	SegER	SegDAER	DAER	SegER
1 (Baseline)	7.9	22.9	29.2	8.1	23.0	29.7
2	7.5	1.9	8.1	7.5	1.9	8.2
3	7.9	2.2	8.5	7.9	2.3	8.6
4	8.2	2.3	8.8	8.3	2.3	8.9

states are final states. These experiments were conducted to determine the effect of the topology of the HMM on the annotation results since previous results showed that the one-state topology produced regular errors in segmentations. The results of these experiments showed that the new topologies provide better segmentation and annotation accuracy in a task-oriented corpus, but that their impact is low (or even negative) in a human-human, non-task-oriented dialogue corpus, whose vocabulary is larger.

Future work is directed towards the integration of the new topologies with other features that showed an increase in the performance of the dialogue annotation, such as the estimation of the number of segments (Tamarit et al., 2009).

Another interesting possibility is to use models with more than one state to obtain the segmentation and then to use the models of one state over the segmented data to obtain the annotation. With this combination, we can avoid the sparseness of the models that are used for the pure annotation, which seems to be the main source of errors in large vocabulary corpora like SwitchBoard. This combination also obtains a better segmentation accuracy as demonstrated by the corpora presented in this study. Other techniques such as the use of the NGT technique presented in (Martínez-Hinarejos, 2009) could also be used to achieve this more accurate segmentation.

Acknowledgements

This work was partially supported by the EC (FEDER/FSE) and the Spanish Government (MEC, MICINN, MITyC, MAEC, “Plan E”), under grants MIPRCV “Consolider Ingenio 2010”-CSD2007-00018, MITTRAL-TIN2009-14633-C03-01, eruditocom-TSI-020110-2009-439, and FPI fellowship BES-2007-16834.

6. References

N. Alcácer, J. Benedí, F. Blat, R. Granell, C. D. Martínez, and F. Torres. 2005. Acquisition and labelling of a

spontaneous speech dialogue corpus. In *Proceedings of 10th International Conference on Speech and Computer (SPECOM)*, pages 583–586, Patras, Greece.

- J. Ang, Y. Liu, and E. Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the ICASSP*, volume 1, pages 1061–1064, Philadelphia.
- 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*, pages 128–139, Málaga, Spain.
- H. Bunt. 1994. Context and dialogue control. *THINK Quarterly*, 3.
- M. G. Core and J. F. Allen. 1997. Coding dialogues with the DAMSL annotation scheme. In David Traum, editor, *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Menlo Park, California. AAAI.
- L. Dybkjær and W. Minker, editors. 2008. *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*. Springer, Dordrecht.
- M. Fraser and G. Gilbert. 1991. Simulating speech systems. *Computer Speech and Language*, 5:81–99.
- T. Fukada, D. Koll, A. Waibel, and K. Tanigaki. 1998. Probabilistic dialogue act extraction for concept-based multilingual translation systems. In *Proceedings of International Conference on Spoken Language Processing*, volume 6, pages 2771–2774.
- 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.
- H. Hardy, T. Strzalkowski, and M. 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.
- C.-D. Martínez-Hinarejos, J.-M. Benedí, and R. Granell. 2008. Statistical framework for a spanish spoken dialogue corpus. *Speech Communication*, 50:992–1008.
- C.-D. Martínez-Hinarejos, V. Tamarit, and J.-M. Benedí. 2009. Improving unsegmented dialogue turns annotation with n-gram transducers. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC23)*, volume 1, pages 345–354, Hong Kong, December. City University of Hong Kong Press.
- C.-D. Martínez-Hinarejos. 2009. A study of a segmentation technique for dialogue act assignation. In *Proceedings of the Eighth International Conference in Computational Semantics IWCS8*, pages 299–304, Tilburg, The Netherlands, January. Tilburg University, Department of Communication and Information Sciences.
- C.-D. Martínez-Hinarejos. 2010. Switchboard corpus preprocessed test version. http://users.dsic.upv.es/~cmartine/research/resources/swbd_preprocesed.tgz.
- 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.
- V. Tamarit, C.-D. Martínez-Hinarejos, and J.-M. Benedí. 2009. Improving unsegmented statistical dialogue act labelling. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing 2009*, pages 434–440, Borovets (BULGARIA), September. RANLP.
- N. Webb and Y. Wilks. 2005. Error analysis of dialogue act classification. In *Proceedings of the 8th International Conference on Text, Speech and Dialogue*, pages 451–458.
- S. Young. 2000. Probabilistic methods in spoken dialogue systems. *Philosophical Trans Royal Society (Series A)*, 358(1769):1389–1402.