Evaluation of different segmentation techniques for dialogue turns

Carlos D. Martínez-Hinarejos, Vicent Tamarit

Departament de Sistemes Informàtics i Computació
Institut Tecnològic d'Informàtica
Universitat Politècnica de València
Camí de Vera, s/n. 46071 València, Spain
{cmartine,vtamarit}@dsic.upv.es

Abstract

In dialogue systems, it is necessary to decode the user input into semantically meaningful units. These semantical units, usually Dialogue Acts (DA), are used by the system to produce the most appropriate response. The user turns can be segmented into utterances, which are meaningful segments from the dialogue viewpoint. In this case, a single DA is associated to each utterance. Many previous works have used DA assignation models on segmented dialogue corpora, but only a few have tried to perform the segmentation and assignation at the same time. The knowledge of the segmentation of turns into utterances is not common in dialogue corpora, and knowing the quality of the segmentations provided by the models that simultaneously perform segmentation and assignation would be interesting. In this work, we evaluate the accuracy of the segmentation offered by this type of model. The evaluation is done on a Spanish dialogue system on a railway information task. The results reveal that one of these techniques provides a high quality segmentation for this corpus.

1. Introduction

Dialogue systems (Kuppevelt and Smith, 2003) are an interesting application of natural language technologies. Generally, in these systems a user asks a computer system for information, and some interaction using dialogue is needed to get the required information. The dialogue strategy is the model that defines the way the system reacts to each user interaction. This strategy could be defined in a rule-based approach (Gorin et al., 1997), but in the last ten years, some data-based strategies have been defined (Young, 2006) for this task. This dialogue strategy usually depends on the interpretation of the user inputs and the previous interactions. The user input must be interpreted in order to obtain the relevant semantics for the dialogue of this input. This semantics is usually coded in the form of Dialogue Acts (DA). A DA labels the intention and function of the corresponding dialogue segment, which is usually known as *utterance* (Stolcke et al., 2000). An utterance is the minimal informational unit from the dialogue viewpoint. Therefore, each dialogue turn may have one or more utterances, and, consequently, many DA can be assigned to a user turn.

Therefore, the assignation of DA to the last user turn is the step that is previous to the answer of the system. Many models have been proposed to perform this assignation (Stolcke et al., 2000), but most of them assume the previous segmentation of the turn before the assignation of the DA. This is clearly an unusual fact, because most dialogue corpora are only turn-segmented. Moreover, in a real implementation of a dialogue system, the user input is not segmented into utterances.

More recently, there have been a few proposals of models that provide both the segmentation and the assignation at the same time (Martínez-Hinarejos et al., 2006). Obviously, this introduces one more difficulty to the problem because the knowledge at the dialogue level is lower with this approach. The results presented in (Martínez-Hinarejos et al., 2006) show that having an accurate segmentation is critical to the good performance of the assignation models.

In this work, we evaluate the segmentation accuracy of two models: one based on Hidden Markov Models and N-grams (HMM-based model), and another based on N-gram Transducers (NGT model). The HMM-based model is the model presented in (Martínez-Hinarejos et al., 2006); however, in this case we are not interested in comparing the performance of the model in the DA assignation task but rather in the segmentation of the turns into utterances. The evaluation is performed on a dialogue corpus that was acquired for the development of a railway information dialogue system in Spanish. This paper is organised as follows: In Section 2. we introduce the models. In Section 3. we present the corpus. In Section 4., we report the experiments and results. In Section 5., we present the conclusions and the future directions in this task.

2. Dialogue Annotation Models

This section presents two dialogue annotation models that have been previously described in other works (Martínez-Hinarejos et al., 2006; Martínez-Hinarejos, 2006). Both models can act on segmented and unsegmented dialogue turns if they are implemented with the appropriate restrictions. In our case, since we are only interested in their performance on segmentation, only the unsegmented application is reported.

2.1. The HMM Model

The problem of assigning a sequence of DA $\mathcal{U} = U_1 U_2 \cdots U_d$ to the sequence of words $\mathcal{W} = w_1 w_2 \cdots w_l$ of a dialogue can be stated as the following optimisation problem:

$$\hat{\mathcal{U}} = \operatorname*{argmax}_{d,\mathcal{U}} \Pr(\mathcal{U}|\mathcal{W})$$

where d is the assigned number of DA. Using the Bayes rule, this optimisation problem can be formulated as:

Work partially supported by VIDI-UPV under project 20070315 and by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01.

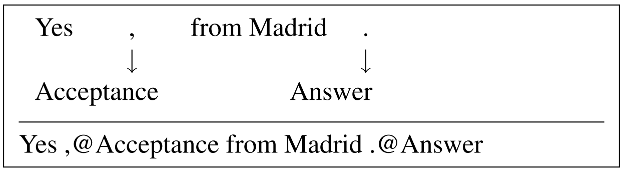

Figure 1: An example of the re-labelling step for dialogue in the GIATI technique.

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

and making the independence assumption, this can be formulated as:

$$\hat{\mathcal{U}} \approx \underset{d,\mathcal{U}}{\operatorname{argmax}} \prod_{i=1}^{d} \Pr(\mathcal{W}_i|U_i) \Pr(U_i|U_1U_2\cdots U_{i-1})$$

where $\mathcal{W}_i$ is the sequence of words that corresponds to DA U_i, restricted to $\mathcal{W}_1\mathcal{W}_2\cdots\mathcal{W}_d = w_1w_2\cdots w_l$.

In this model, $\Pr(\mathcal{W}_i|U_i)$ is modelled by a Hidden Markov Model (HMM) whose emitted symbols are the words of the dialogue, and $\Pr(U_i|U_1U_2\cdots U_{i-1})$ is modelled by an N-gram (i.e., $\Pr(U_i|U_1U_2\cdots U_{i-1}) \approx \Pr(U_i|U_{i-1-n}\cdots U_{i-1})$). Each model can be estimated from labelled data using the classical Baum-Welch algorithm for HMM and the N-gram estimation algorithms for the N-gram.

This model can be applied in a turn-by-turn fashion (which is closer to its use in a real implementation of a dialogue system) or in a complete-dialogue fashion (which is more appropriate for the annotation of transcribed dialogues). The search problem can be solved with the classical Viterbi algorithm, which produces the segmentation of the turns into utterances as a by-product ($\mathcal{W}_i$ in the used notation). More details on this model and its implementation can be found in (Martínez-Hinarejos et al., 2006).

2.2. The NGT Model

The NGT model is based on a Finite-State Transducer (FST) inference technique that is known as GIATI (Casacuberta et al., 2005). This technique was initially proposed to solve general machine translation problems, but it can be adapted to the dialogue problem.

The GIATI process starts from a parallel and aligned corpus of input-output sentences. Based on this alignment, a re-labelling step is applied to produce a set of strings in a new language. An N-gram is inferred from these strings, and then an inverse re-labelling step is applied to convert the N-gram into an FST. This last step is difficult when smoothing techniques are applied in the N-gram inference. To skip this step, a Viterbi decoding on N-grams has been implemented.

In the case of the dialogue problem, the words of the turns are considered as the input language, and the DA labels constitute the output language. The re-labelling step attaches the DA label of a segment to the last word of that segment. An example of this re-labelling step is provided in Figure 1. The Viterbi decoding takes the unlabelled words and processes them in a tree-search, to obtain an annotation of the dialogue turns.

This model can be applied in the turn-by-turn and complete-dialogue fashions as well. As in the HMM-based model, the Viterbi decoding produces the segmentation into utterances as a by-product. More details on this model and its implementation can be found in (Martínez-Hinarejos, 2006).

3. The Dihana Corpus

The experiments carried out in this work were performed on the dialogue corpus acquired for the DIHANA project (Benedí et al., 2004). The goal of the DIHANA project was the construction of a modular speech dialogue system, which is devoted to the consultation of timetables and fares for Spanish trains nationwide.

The dialogue corpus was acquired using the Wizard of Oz (WoZ) (Fraser and Gilbert, 1991) set-up. The users were required to accomplish four different scenarios of different nature, without lexical or syntactical restrictions (spontaneous speech). A total number of 900 dialogues were acquired for the project, with a total of more than 15,000 turns (6,280 for user and 9,133 for system). These dialogues were manually transcribed and annotated at the dialogue level.

Each label is defined by three different levels. The first level corresponds to the speech act (intention) of the annotated utterance, and it is task-independent. The second and third levels correspond to the data repository (frame) used in the utterance and the specific data (cases) given in the utterance. The second and third levels are task-dependent. In this case, they cope with the classical concepts which are managed in a railway environment (departure and destination towns, times, fares, dates, etc.). Table 1 shows the possible values for each level. The first level admits only one value, but the second and third levels may be multivaluated.

The set of DA labels was composed of 248 different labels (153 for user and 95 for system). The resulting number of utterances were 9,712 for user turns and 13,830 for system turns. More details can be found in (Alcácer et al., 2005).

To reduce the number of labels, an alternative labelling was produced using only the first two levels of the labels. With this reduction, the total number of different labels was 72 (45 for user and 27 for system). The resulting number of utterances with this alternative annotation scheme was 7,014 for user turns and 13,828 for system turns.

The final vocabulary was composed of 980 words. Before applying the previously described annotation techniques, some preprocessing steps were performed: categorisation (e.g., town names, hours, dates, ...) and identification of the speaker for each word (i.e., to identify if the word corresponded to a user or a system turn). With these preprocessing steps, the total number of different words in the vocabulary was 895.

4. Experiments and Results

The experiments were directed to the evaluation of the accuracy of the segmentation given by the models. All the experiments were performed turn-by-turn in the application

First level	Second level	Third level
Opening, Closing, Undefined, Not-Understood, Waiting, Consult, Acceptance, Rejection, Question, Confirmation, Answer	Nil, Dep-time, Arr-time, Fare, Org, Dest, Day, Train-type, Service, Class, Trip-time	Nil, Dep-time, Arr-time, Fare, Org, Dest, Day, Train-type, Service, Class, Trip-time, Order-num, Num-trains, Trip-type

Table 1: Labels defined for each level. The *Nil* value denotes the absence of information.

N-gram	2 levels		3 levels	
	WER	SER	WER	SER
2	11.3%	12.2%	38.1%	34.3%
3	11.3%	12.2%	38.4%	34.7%
4	11.4%	12.1%	38.1%	34.5%

Table 2: Transcription experiments (five-fold cross-validation) for the HMM-based model.

of the models, which is closer to the use of the models in a real dialogue system. The evaluation was only performed on user turns because system turns are not to be recognised in a dialogue system.

Two types of experiments were performed to assess the models:

Transcription experiments These experiments used the transcribed dialogues, which are supposed to be free of recognition errors (i.e., we simulated a perfect recognition system). A cross-validation approach was used in this case. Five different partitions of 180 dialogues were constructed, using four partitions for the training of the models and the remaining one for testing. The input for the models were the manual transcriptions of the dialogues.

Recognition experiments One of the previously defined partitions was recognised with a continuous speech recogniser (20% of WER), eliminating punctuation marks. A segmentation of the recognised turns was performed based on the segmentation of the real transcription. In this case, the speech signal was aligned with the real transcription and the signal segments provided the utterances of the categorised recognised sentence. The other four partitions were used for training purposes removing the punctuation marks but keeping the categorisation.

In the case of the HMM-based model, a fixed weight factor (which offered the best results in previous experiments) was used to balance the influence of the language model. The assessment was done with two error measures: WER (which accounts for partially correct segmentations) and SER (which account for incorrect segmentations). The results for the transcription experiments are presented in Tables 2 and 3. The results for the recognition experiment are presented in Tables 4 and 5.

These results clearly show that the NGT model produces better segmentations than the HMM-based model in the general case. Moreover, the more precise the labelling scheme is, the better the improvement of the NGT model with respect to the HMM-based model. However, from the results in Tables 4 and 5, we can infer that the NGT model is more sensitive to the quality of the recognition, and specially sensitive to the lack of punctuation marks (which are usually attached to the DA labels in the re-labelling process). In any case, the NGT model performed better than the HMM-based model in the finer labelling (3-level labelling).

A more detailed error analysis was done on the best results for each model and labelling scheme. With respect to the number of utterances each technique produced, the HMM-based model used to obtain the same number of utterances in the incorrectly segmented turns; this ocurred in 71% of the turns in the 2-level labelling scheme and in 50% of the turns in the 3-level labelling scheme. However, the erroneous segmentations of the NGT model usually had a different number of utterances from the reference segmentation; only 2.5% of the turns with wrong segmentation in the 2-level labelling scheme and 9% in the 3-level labelling

N-gram	2 levels		3 levels	
	WER	SER	WER	SER
2	9.1%	9.8%	10.3%	13.8%
3	6.3%	6.6%	8.3%	11.2%
4	6.2%	6.5%	7.8%	10.6%
5	6.9%	7.1%	7.8%	10.7%

Table 3: Transcription experiments (five-fold cross-validation) for the NGT model.

N-gram	2 levels		3 levels	
	WER	SER	WER	SER
2	10.4%	11.1%	36.5%	33.3%
3	10.8%	11.4%	37.3%	33.7%
4	10.6%	11.3%	36.8%	33.7%

Table 4: Recognition experiments for the HMM-based model.

N-gram	2 levels		3 levels	
	WER	SER	WER	SER
2	20.4%	20.4%	26.5%	30.3%
3	23.0%	22.4%	29.0%	32.2%
4	27.4%	26.1%	30.9%	34.1%
5	29.1%	27.3%	31.0%	34.0%

Table 5: Recognition experiments for the NGT model.

had the same number of utterances than the reference segmentation. Although these results were obtained for the transcription experiments, similar results were obtained for the recognition experiments. This reveals the very different nature of the errors produced by the two techniques.

With respect to the nature of the wrongly segmented turns, in the case of the 3-level labelling scheme, the dispersion of the different errors was very high and no real comparison of the nature of the errors could be carried out. In the case of the 2-level labelling scheme, the nature of the erroneously labelled turns was very different for the two models. The HMM-based model presented a higher number of errors in the case of turns with two utterances, where the first utterance was an *acceptance* or *rejection* and the second utterance was an *answer* about *departure hours*. However, the NGT model produced more errors in the case of turns of only one utterance which was labelled with a *question* about *departure time* or *fares*. Both models produced a similar amount of errors in two types of two-utterance turns: *answers* about *day* with *answers* about *departure hour*, *acceptances* about *departure hour* with *answers* about *departure hour*.

This analysis of the different nature of the errors produced offers the possibility of combining the two models in order to obtain a better segmentation of the dialogue turns.

5. Conclusions and Future Work

From the results, we can conclude that, in general, the NGT model has better behaviour in the segmentation process than the HMM-based model. However, previous works have demonstrated that the NGT model has lower accuracy in the assignation of DA to user turns. Therefore, clearly the next step is the combination of these two models into a two-stage DA assignment: a first segmentation stage using the NGT model and a second assignation stage using the HMM-based model.

Another possible work is the definition of generic segmentation models, which are independent from the DA labels of the training corpus. This can be achieved easily with the NGT model by using a unique DA label. Another line of research to explore is the direct combination of the two models to obtain more accurate segmentations.

As the 3-level segmentation is more difficult to achieve, one possible solution is to perform the 2-level segmentation and to reapply the segmentation models over each of the 2-level utterances to obtain the 3-level utterances. Since the quality of the two-level segmentation is high, this would ensure that more segmentation points are adequately positioned.

In any case, this is an initial work and these conclusions should be confirmed with more extensive experiments on other corpora like SwitchBoard (Godfrey et al., 1992) or CallHome (Levin et al., 1999).

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 *Proceeding of 10th International Conference on Speech and Computer (SPECOM)*, pages 583–586, Patras, Greece.

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.

F. Casacuberta, E. Vidal, and D. Picó. 2005. Inference of finite-state transducers from regular languages. *Pattern Recognition*, 38(9):1431–1443.

M. Fraser and G. Gilbert. 1991. Simulating speech systems. *Computer Speech and Language*, 5:81–99.

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

A. L. Gorin, G. Riccardi, and J. H. Wright. 1997. How may I help you? *Speech Communication*, 23(1/2):113–127.

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.

Lori Levin, Klaus Ries, Ann Thymé-Gobbel, and Alon Levie. 1999. Tagging of speech acts and dialogue games in Spanish call home. In Marilyn Walker, editor, *Towards Standards and Tools for Discourse Tagging: Proceedings of the Workshop*, pages 42–47. Association for Computational Linguistics, Somerset, New Jersey.

C. D. Martínez-Hinarejos, R. Granell, and J. M. Benedí. 2006. Segmented and unsegmented dialogue-act annotation with statistical dialogue models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sesions*, pages 563–570, Sydney, Australia,, 17th-21th July.

C.D. Martínez-Hinarejos. 2006. Automatic annotation of dialogues using n-grams. In Petr Sojka, Ivan Kopeček, and Karel Pala, editors, *Proceedings of the Ninth International Conference on Text, Speech and Dialogue—TSD 2006*, Lecture Notes in Artificial Intelligence LNCS/LNAI 4188, pages 653–660, Brno, Czech Republic, Sep. Springer-Verlag.

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.

Steve Young. 2006. Using pomdps for dialog management. In *Proc. IEEE/ACL Workshop on Spoken Language Technology (SLT 2006)*, pages 8–13.