Estimating the Number of Segments of a Turn in Dialogue Systems*

Vicent Tamarit and Carlos-D. Martínez-Hinarejos

Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera s/n
46022 Valencia, Spain

Abstract. An important part of a dialogue system is the correct labelling of turns with dialogue-related meaning. This meaning is usually represented by dialogue acts, which give the system semantic information about user intentions. This labelling is usually done in two steps, dividing the turn into segments, and classifying them into DAs. Some works have shown that the segmentation step can be improved by knowing the correct number of segments in the turn before the segmentation. We present an estimation of the probability of the number of segments in the turn. We propose and evaluate some features to estimate the probability of the number of segments based on the transcription of the turn. The experiments include the SwitchBoard and the Dihana corpus and show that this method estimates correctly the number of segments of the 72% and the 78% of the turns in the SwitchBoard corpus and the Dihana corpus respectively.

1 Introduction

A dialogue system is usually defined as a computer system that interacts with a human user to achieve a task using dialogue [5]. The computer system must interpret the user input, in order to obtain the meaning and the intention of the user turn. This is needed to give the appropriate answer to the user. The selection of this answer, along with other decisions that the system can take, is guided by the so-called dialogue strategy. This dialogue strategy can be rule-based [9] or data-based [17].

In either case, the dialogue strategy needs the interpretation of user turns to achieve the aim of the user. This interpretation must only take into account the essential information for the dialogue process, which is usually represented by special labels called Dialogue Acts (DA) [4]. With this approximation, each user turn can be divided into non-overlapped sequences of words, and each sequence is classified into the available DAs. These sequences of words are usually called segments (some authors call them utterances [15]). Each segment has an associated class (DA) which defines its dialogue-related meaning (usually the intention, the communicative function, and the important data).

* Work supported by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01 and by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

Tamarit V. and Martínez-Hinarejos C. (2009).
Estimating the Number of Segments of a Turn in Dialogue Systems.
In *Proceedings of the 9th International Workshop on Pattern Recognition in Information Systems*, pages 9-17
DOI: 10.5220/0002173600090017

In recent years, probabilistic data-based models have gained importance for this task, such as decision trees or neural networks [16]. The dialogue corpora provide sets of dialogues that are divided into segments and annotated with DA labels. These dialogues are the data used to estimate the probabilistic parameters of the data-based models. This model usually contains two modules: the segmentation module, which estimates the segments of the turn, and the classification module, which classifies the segment. In the posterior use of the models, they are applied to non-annotated dialogues to divide the turn and obtain the most likely DA for each segment.

Most of the previous work on DA assignation assumed the correct segmentation of the dialogue, so the problem is reduced to a classification task [12]. However, in a real situation, the only data that are available are the dialogue turns. The models can be adapted to the real situation in which segmentation is not available, but, in this case, the labelling accuracy is lower than that produced over correctly-segmented dialogue turns [14].

Some authors proposed obtaining a segmentation hypothesis from some lexical and prosodic features [2]. The work presented good results but the classification task is limited to 5 classes and is oriented only to spoken dialogs.

Instead of estimating the entire segmentation, another less restricting possibility is to estimate the number of segments of a given turn. Once the estimation is made, the search for the most likely DA sequence is restricted to only having the estimated number of DA. The estimation of the number of segments can be done using the transcriptions of the turns, so it is possible to use it in typed dialogues, where only the text is available, and in spoken dialogues.

Some works [13] have shown that the labelling is improved when there is a correct estimation of the number of segments of a turn. In this paper, we present a model to estimate the number of segments given the transcription of the turns, and using turn transcription derived features for the estimation.

The paper is organised as follows: In Section 2 we introduce the model proposed for the estimation of the number of segments along with the used features. In Section 3 we present the corpora used to test the method, and the results of the performed experiments. In Section 4 we present our final conclusions and future work.

2 Estimation of the Number of Segments

Given a word sequence of l words $W = w_1 w_2 \dots w_l$, we define the probability for a turn W to have r segments as $\Pr(r|W)$. We approximate this probability as $\Pr(r|S_c)$, where S_c is a score based on the sequence of words ($S_c = f(W)$).

The probability of r, given the score, can be calculated by applying the Bayes rule:

$$\Pr(r|S_c) = \frac{p(S_c|r)p(r)}{p(S_c)} \tag{1}$$

The a priori probability $p(r)$ can be easily computed as the number of utterances with r segments, N_{Tr}, divided by the total number of turns N_T:

$$p(r) = \frac{N_{Tr}}{N_T} \tag{2}$$

The conditional member $p(S_c|r)$ is estimated by a normal distribution. We calculated one distribution for each r:

$$p(S_c|r) \sim \mathcal{N}(m_r, \sigma_r) \tag{3}$$

The mean m_r and standard deviation σ_r are computed from the scores associated with the turns with r segments.

The last element $P(S_c)$ is estimated by another gaussian distribution that is computed from all the turns:

$$p(S_c) \sim \mathcal{N}(m_{Sc}, \sigma_{Sc}) \tag{4}$$

The mean m_{Sc} and standard deviation σ_{Sc} are computed from the all the scores in the training data.

2.1 Possible Features

The computation of S_c is made using features that are extracted from the transcription of each turn (it is word-based). One evident feature is the number of words of the turn. More sophisticated features can be inferred from the words (or sequences) that usually appear at the beginning or the end of segments. We made a study of the features that could determine the number of segments and we evaluated the influence of some of them:

- Length of the turn. We evaluated the relation between the number of segments and the number of words in a turn.
- Final words and n-grams. In the transcription, some words (like the interrogation mark and the period) clearly indicate the end of a segment. Combinations of the last two or three words are also useful.
- Initial words and n-grams. This is the opposite case to the final words and n-grams.
- Combinations: The above features can be combined to obtain a better estimation of the number of segments.

2.2 Basic Scores

Second, we defined some calculations for the score S_c based on the above-mentioned features. This scores use only one of the proposed features.

- Based on length of the turn
  The score S_c can be calculated as the number of words in the turn:

$$S_c(W) = l \tag{5}$$

– Boundary words
We define the score S_c of a turn W as:

$$S_c(W) = \sum_{i=1}^{l} p_f(w_i) \tag{6}$$

where $p_f(w_i)$ is the probability of the word i being a final word in a segment. It is estimated by counting the number of times that the word is final divided by the total number of appearances of the word. This value is 0 for the words that never appear at the end of a segment.
It is also possible to calculate S_c in the same way using the initial words of a segment instead of the final ones.
– Boundary n-grams
Instead of calculating the probability of a final word, we propose the estimation of the probability of the n last words of the segments. In this case, the method of estimation is the same one that we used in the above case: the number of times that the n-gram is at the end of the segment divided by the total number of appearances of the n-gram. We calculated the S_c using that estimation with:

$$S_c(W) = \sum_{i=n}^{l} p_f(w_{i-(n-1)}^{i}) \tag{7}$$

As we proposed in the final word estimation, the probability of initial n-grams in a segment can be computed just by counting the times an n-gram is initial.

The features that we used in the estimation of the score can be combined in two different ways: composing a score from different features or by a naive-Bayes computation. We explore these possibilities in the following subsections

2.3 Composed Score

In this combined form, the calculated score for a turn is composed of various features, e.g. the score can be seen as the summation of the probability of each word to be final plus the length of the turn:

$$S_c(W) = l + \sum_{i=1}^{l} p_f(w_i) \tag{8}$$

Another option is to combine the final words with final n-grams, e.g., combining the final bigrams and the final words:

$$S_c(W) = \sum_{i=2}^{l} p_f(w_{i-1}^{i}) + \sum_{i=1}^{l} p_f(w_i) \tag{9}$$

Using this method, we can combine any of the basic features.

2.4 Naive-Bayes Computation

In the naive-Bayes computation, the final probability of the number of segments is calculated by combining the probabilities for each score, i.e., if we consider:

$$\Pr(r|S_{c_1}, S_{c_2}, \cdots S_{c_n}) \tag{10}$$

this probability can be simplified assuming that there are no dependencies between scores (naive-Bayes assumption):

$$\begin{aligned} &\Pr(r|S_{c_1}, S_{c_2}, \cdots S_{c_n}) = \\ &\Pr(r|S_{c_1}) \Pr(r|S_{c_2}) \cdots \Pr(r|S_{c_n}) \end{aligned} \tag{11}$$

3 Experiments and Results

We present a set of experiments that we performed using the SwitchBoard corpus [8] and the Dihana corpus [3]. The experiments were designed to show the error in the estimation of the number of segments using the estimation proposed in Section 2. We compared the different methods and two versions of the corpora: one version contains the correct transcriptions, which include all the punctuation marks, and exclamation/interrogation marks, and the other version does not include those symbols, and it is included as an approximation to the output of a (perfect) speech recogniser.

3.1 SwitchBoard Corpus

The SwitchBoard corpus is a well-known corpus of human-human conversations by telephone. The conversations are not related to a specific task, since the speakers discuss general interest topics, with no clear task to accomplish. This corpus recorded spontaneous speech, with frequent interruptions between the speakers and background noises. The transcription of the corpus takes into account all these facts and it includes special notation for the overlaps, noises and other sound effects present in the acquisition.

The corpus is composed of 1,155 different conversations in which 500 different speakers participated. The number of turns in the dialogues is around 115,000, including overlaps. The vocabulary size is approximately 42,000 words.

The corpus was manually divided into segments following the criteria defined by the SWBD-DAMSL annotation scheme [10]. Each segment is labelled with one of the 42 different labels present in the SWBD-DAMSL annotation set. These labels represent categories such as statement, backchannel, questions, answers, etc., and different subcategories for each of these categories (e.g., statement opinion/non-opinion, yes-no/open/rethorical-questions, etc.). The manual labelling was performed by 8 different human labellers, with a Kappa value of 0.80.

To simplify the task, we preprocessed the transcriptions of the SwitchBoard corpus to remove certain particularities. The interrupted turns were joined, thereby avoiding interruptions and ignoring overlaps between the speakers. The vocabulary was reduced

by using all the words in lowercase and separating the punctuation marks from the words.

To obtain more reliable results, we performed a partition on the corpus to perform experiments with a cross-validation approach. In our case, the 1,155 different dialogues were divided into 11 partitions with 105 dialogues each one.

3.2 Dihana Corpus

The Spanish corpus Dihana [3] is composed of 900 dialogs about a telephonic train information system. It was acquired by 225 different speakers (153 male and 72 females), with small dialectal variants. There are 6,280 user turns and 9,133 system turns. The vocabulary size is 823 words. The total amount of speech signal was about five and a half hours.

The acquisition of the Dihana corpus was carried out by means of an initial prototype, using the Wizard of Oz (WoZ) technique [6]. This acquisition was only restricted at the semantic level (i.e., the acquired dialogues are related to a specific task domain) and was not restricted at the lexical and syntactical level (spontaneous-speech). In this acquisition process, the semantic control was provided by the definition of scenarios that the user had to accomplish and by the WoZ strategy, which defines the behaviour of the acquisition system.

The annotation scheme used in the corpus is based on the Interchange Format (IF) defined in the C-STAR project [11]. Although it was defined for a Machine Translation task, it has been adapted to dialogue annotation [7]. The three-level proposal of the IF format covers the speech act, the concept, and the argument, which makes it appropriate for its use in task-oriented dialogues.

Based on the IF format, a three-level annotation scheme of the Dihana corpus segments was defined in [1]. This DA set represents the general purpose of the segment (first level), as well as more precise semantic information that is specific to each task (second and third levels).

All of the dialogues are segmented in turns (User and System), and each turn is also divided into segments. Finally, each segment is labelled with a three-level label. Obviously, more than one segment can appear per turn. In fact, an average of 1.5 segments per turn was obtained.

The corpus is divided into 5 partitions, so the experiments can be performed with a cross-validation approach. Each partition contains 180 dialogues. In this work we only used the user turns.

3.3 Estimation of the Number of Segments

We used the method proposed in Section 2 to estimate the number of segments of the turns in the SwitchBoard and Dihana corpora. We did three different subsets of experiments. The first subset includes estimation of segments with simple scores, the second subset refers to the estimation for the composed scores, and the third subset tests the naive-Bayes computation of the score.

The results are presented in tables 1 and 2. Table 1 shows the results of the different estimations of the number of segments for both corpora using the basic scores presented

Table 1. Results of the estimation of the number of segments using basic scores. The estimation column indicates the type of the score used in the estimation of r. The error indicates the percentage of turns with an estimation of the wrong number of segments using the SwitchBoard and Dihana corpora. We included two versions of the corpora, the real transcription and a filtered one, without interrogation and punctuation marks. Best results for each corpus are shown in boldface.

	Simple score			
Estimation	SWBD	SWBD no marks	Dihana	Dihana no marks
Length	35.8	36.0	30.9	30.8
Final Words	33.4	34.7	25.1	25.7
Final Bigrams	**28.1**	33.6	**22.2**	**25.5**
Final Trigrams	39.1	39.1	31.9	32.2
Initial Words	33.4	**33.4**	33.1	33.3
Initial Bigrams	32.6	34.6	29.1	30.9
Initial Trigrams	39.0	37.6	33.0	32.3

Table 2. Results of the estimation of the number of segments using composed scores and estimations with the naive-Bayes approach. The error indicates the percentage of turns with an estimation of the wrong number of segments using the SwitchBoard and Dihana corpora. Best results for each corpus are shown in boldface.

	Composed features			
Estimation	SWBD	SWBD no marks	Dihana	Dihana no marks
Length + Final Words	35.6	35.9	30.4	30.0
Length + Final Bigrams	35.6	35.9	30.6	30.0
Length + Initial Words	35.6	35.7	30.9	30.5
Length + Initial Bigrams	35.5	35.8	30.7	30.1
Final Bigrams + Initial Words	30.3	**32.5**	27.1	28.1
Final Bigrams + Initial Bigrams	**29.0**	33.5	**23.2**	**27.2**

	Naive-Bayes computation			
Estimation	SWBD	SWBD no marks	Dihana	Dihana no marks
Length + Final Words	34.4	35.3	28.3	26.9
Length + Final Bigrams	32.9	35.0	22.7	**26.7**
Length + Initial Words	34.5	34.5	30.6	30.5
Length + Initial Bigrams	33.6	34.8	28.0	29.3
Final Bigrams + Initial Words	30.2	**33.4**	22.7	27.4
Final Bigrams + Initial Bigrams	**29.5**	33.9	**22.4**	27.3

in subsection 2.2. Table 2 includes the results of the composed score presented in subsection 2.3, and the estimation with the naive-Bayes approach presented in subsection 2.4.

In all the tables, we included the estimation using the correct turn transcriptions and an approximation to the output of a speech recogniser, where some marks are not present. Specifically, we deleted the exclamation and interrogation marks and the punctuation marks.

These tests showed that the models work in the same way for both corpora. The final bigrams seem to be the best feature in Dihana. In Switchboard, the error with final

bigrams using the no-marks transcription is increased by a 20% with respect to the correct transcription. In both corpora, the initial n-grams produce worse results than the final ones.

The final word and trigram features produce similar errors in both versions of the corpora, which indicates that the marks we deleted are not as useful as can be expected for determining the number of segments. Nevertheless, when we use the final bigram feature, the absence of punctuation marks slightly affects the estimation of the number of segments. The length of the turn showed no important variation between the two versions of the corpora.

The two proposed ways to do the combination of scores produces similar optimal estimations. Nevertheless, in both corpora the naive-Bayes computation produces better results on average than the composed score. In Dihana, the combined scores in which the length of the turn is present, work better with the no-marks version of the corpus. SwitchBoard does not present important differences in this case.

In the two corpora, the best results for both approximations of composed scores show that the final bigram is a good estimator, because the combined scores where this feature is present produce the best estimations. In the Dihana corpus, the composed estimations do not produce any improvement in the estimation of the number of segments. However, in the Switchboard corpus the best estimation of the number of segments is produced by a composed score from final bigrams and initial words.

4 Conclusions and Future Work

In this work, we proposed a method to estimate the number of segments of a turn given its transcription. This method can use some transcription derived features, so it can be used in spoken or typed dialogues. We compared the different features presented using two corpora. The experiments showed that the punctuation marks are not essential to identify segments in a turn. Moreover, the composed estimations seem to produce good results even without marks.

Future work is directed to obtaining a better model that estimates the number of segments. For example, the model can combine more than two features and use some sort of weights for the features. Another important step is to introduce the estimation of the number of segments in a labelling model. However, the estimations based on the transcription of turns do not seem to produce good enough results. In spoken dialogues, a new estimation could be to use features that are extracted directly from the audio signal, as proposed in [2]. Therefore, studying the audio features and including them into our probability model of the estimation of the number of segments could be a good idea.

References

1. N. Alcácer, J. Benedí, F. Blat, R. Granell, C. D. Martínez, and F. Torres. Acquisition and labelling of a spontaneous speech dialogue corpus. *Proceeding of 10th International Conference on Speech and Computer (SPECOM). Patras, Greece*, pages 583–586, 2005.

2. J. Ang, Y. Liu, and E. Shriberg. 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, 2005.
3. J.-M. Benedí, E. Lleida, A. Varona, M.-J. Castro, I. Galiano, R. Justo, I. López de Letona, and A. Miguel. Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana. *Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1636–1639, May 2006.
4. H. Bunt. Context and dialogue control. *THINK Quarterly*, 3, 1994.
5. L. Dybkjaer and W. Minker. *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*. Springer, 2008.
6. M. Fraser and G. Gilbert. Simulating speech systems. *Computer Speech and Language*, (5):81–89, 1991.
7. T. Fukada, D. Koll, A. Waibel, and K. Tanigaki. Probabilistic dialogue act extraction for concept based multilingual translation systems. *ICSLP 98*, pages 2771–2774, 1998.
8. J.J. Godfrey, E.C. Holliman, and J. McDaniel. Switchboard: telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1:517–520, 1992.
9. A. Gorin, G. Riccardi, and J. Wright. How may I help you? *Speech Communication*, 23:113–127, 1997.
10. D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard swbd-damsl shallow- discourse-function annotation coders manual - draft 13. Technical Report 97-01, University of Colorado Institute of Cognitive Science, 1997.
11. A. Lavie, L. Levin, P. Zhan, M. Taboada, D. Gates, M. Lapata, C. Clark, M. Broadhead, and A. Waibel. Expanding the domain of a multi-lingual speech-to-speech translation system. *Proceedings of the Workshop on Spoken Language Translation, ACL/EACL-97*, 1997.
12. L. Levin, K. Ries, A. Thymé-Gobbel, and A. Levie. Tagging of speech acts and dialogue games in Spanish call home. In *the Workshop: Towards Standards and Tools for Discourse Tagging*, pages 42–47, 1999.
13. C.D. Martínez-Hinarejos. 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 Nederlands, January 2009. Tilburg University, Department of Communication and Information Sciences.
14. C.D. Martinez-Hinarejos, J.M. Benedi, and R. Granell. Statistical framework for a spanish spoken dialogue corpus. *Speech Communication*, 50:992–1008, 2008.
15. A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):1–34, 2000.
16. D. Vilar, M.J. Castro, and E. Sanchis. Connectionist classification and specific stochastic models in the understanding process of a dialogue system. In *European Conference on Speech Communication and Technology*, pages 645–648, Geneva, Switzerland, September 2003.
17. S. Young. Probabilistic methods in spoken dialogue systems. *Philosophical Trans Royal Society (Series A)*, 358(1769):1389–1402, 2000.