On the training data requirements for an automatic dialogue annotation technique*

Carlos D. Martínez-Hinarejos
Departamento de Sistemas Informáticos y Computación
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia, Cno. de Vera s/n, Valencia, 46022
cmartine@dsic.upv.es

Abstract

When constructing a task-oriented dialogue system, it is usual to perform an acquisition of dialogues for the system's task. This acquisition can be used to define the behaviour of the dialogue system, and it can be rule-based or corpus-based. In the corpus-based case, the models that define the behaviour are automatically inferred from annotated dialogues. The annotation process is time-consuming and error-prone, and the use of assistant tools for the annotation can reduce the effort in this process. In this work, the data requirements of a previously presented annotation tool are presented, and the results show that the technique obtains its maximum performance even with a relative small amount of annotated dialogues.

1 Introduction

A dialogue system (Kuppevelt and Smith, 2003) is usually defined as an automatic system that interacts with a human user using dialogue, with the objective of solving a certain problem. Tasks such as timetable consultation (Aust et al., 1995) are common examples of dialogue system applications.

A dialogue system is defined by its behaviour, which tries to imitate a real dialogue situation. The most common method to define this behaviour is to acquire a corpus of dialogues on the task to be solved. In this acquisition, a set-up known as Wizard of Oz (Fraser and Gilbert, 1991) is used.

Then, the behaviour of the system is defined by analysing the acquired corpus of dialogues. Two main approximations have been used in the system's behaviour definition: rule-based (Gorin et al., 1997) and corpus-based (Stolcke et al., 2000). In the corpus-based approach the behaviour is determined by statistical models that are automatically inferred and updated. Therefore, in the corpus-based approach it is easier to adapt the system behaviour to new tasks and situations by inferring a new model.

The corpus-based approach needs huge amounts of data (dialogues) conveniently annotated to estimate the parameters of the statistical models. The most widely used annotation scheme is the Dialogue Act (DA) labelling (Searle, 1969). In this scheme, every turn of the dialogue is segmented into a utterance (Stolcke et al., 2000) and annotated with one DA, which defines its function in the dialogue.

The annotation of the corpus implies the definition of the set of DA and the annotation rules (Alcacer et al., 2005; Jurafsky et al., 1997), followed by the annotation itself, which is a very time-consuming process. Therefore, the development of automatic annotation techniques is very useful in the development of corpus-based dialogue systems.

Some automatic annotation techniques have been proposed in previous works (Stolcke et al., 2000). These techniques use part of the annotated dialogue corpus to infer the automatic annotators. These annotators are statistical models that, given the sequence of words, return the utterances with their corresponding DA labels. The automatic annotators are not error-free, and they improve their error rate as long as more training data is provided.

In this work, the influence of the amount of training dialogues on the automatic annotator error rate is presented. The results show that when using more than a certain number of dialogues, no significant improvements in the annotation error rate are no-

*Work partially supported by VIDI-UPV under PAID06-20070315 program.

Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 211–214,
Antwerp, September 2007.

ticed. This allows to determine the size of the corpus that must be manually annotated to obtain the highest automatic annotation performance with the lowest manual annotation cost.

The paper is organised as follows. In Section 2, the annotation technique is presented. In Section 3, the used dialogue corpora are described. In Section 4, the performed experiments and their results are discussed. In Section 5, conclusions and future work lines are presented.

2 GIATI based annotation technique

The automatic annotation technique which is analysed in this work is based on a general Stochastic Finite-State Transducer (SFST) inference technique known as GIATI (Casacuberta et al., 2005). This technique has been successfully used in Machine Translation tasks and in dialogue annotation (Martínez-Hinarejos, 2006).

GIATI infers a SFST from a parallel corpus using a re-labelling process of input-output pairs of sentences. From the re-labelled corpus, a smoothed n-gram is inferred and then it is converted into the final SFST by reverting the initial re-labelling. A modification of the GIATI annotation was proposed in (Martínez-Hinarejos, 2006) to perform the annotation directly using the n-gram instead of the SFST.

In dialogue is easy to find a re-labelling scheme because no cross-alignments are usually present (a DA label is attached to a complete utterance in a linear manner). For example, the DA label can be attached to all the words in the corresponding utterance, or only to the last word of the utterances. In this work, this last re-labelling strategy is used, following the steps presented in (Martínez-Hinarejos, 2006).

After the inference from the re-labelled corpus, the n-gram can be used as an annotator model. For the annotation, a Viterbi n-gram implementation was used following the ideas of (Martínez-Hinarejos, 2006). Intensive beam-search was applied in the implementation to avoid the problems with large exploration trees in the Viterbi process.

3 Dialogue corpora

In the experiments, two different dialogue corpora, with very different features, were used to assess the performance of the automatic annotation technique.

3.1 Dihana corpus

Dihana (Benedí et al., 2004) is a task-oriented corpus which is composed of computer-to-human dialogues. The main aim of the task is to answer telephone queries about timetables, fares, and services for long distance trains. The language of the corpus is Spanish.

The corpus is composed of 900 different dialogues that were acquired using the Wizard of Oz technique and semicontrolled scenarios. The total set of dialogues comprises 6,280 user turns and 9,133 system turns, with a vocabulary of 980 words. All the dialogues were annotated by human experts. The annotation scheme used in Dihana was presented in (Alcacer et al., 2005). The labels are organised in three different levels. The total number of labels which are present in the corpus is 248 (153 for user turns and 95 for system turns). If only the first and second level are taken into account, 72 different labels (45 for user and 27 for system) are present.

3.2 SwitchBoard corpus

SwitchBoard (Godfrey et al., 1992) is a well-known speech corpus which was obtained from human-to-human telephone conversations. These conversations were not task oriented, and both speakers were allowed to express themselves in a free manner and to interrupt the other speaker, discussing a general topic, but with no task to accomplish.

The corpus is composed of 1,155 conversations, with a total number of 126,754 different turns of spontaneous speech. The vocabulary size is 42,672 words. The corpus was annotated using a simplification of the DAMSL annotation scheme (Jurafsky et al., 1997) which comprises a total number of 42 different labels.

4 Experiments and results

The objective of the experiments is to determine which amount of labelled dialogues is enough to obtain the best possible GIATI-based dialogue labeller with the minimum annotation effort. It is clear that the quality of the labellers should be assessed with respect to a set of dialogues which is not included in the training corpus and that it must be fixed for all the variable-size training corpora. In our case, a set of 100 dialogues of the Dihana corpus and a set of 155 dialogues of the SwitchBoard corpus were taken

as the test corpora. The sizes of the training corpora were from 100 up to 800 in the Dihana corpus, and up to 1,000 in the case of the SwitchBoard corpus (with increments of 100 dialogues).

Some common preprocessing steps were performed in order to reduce data sparseness: case unification (all the words were transcribed in lowercase) and punctuation marks treatment (the punctuation marks were separated from the words).

For the Dihana corpus, two more preprocessing steps were applied: a categorisation (it was performed for categories such as town names, the time, dates, etc.) and the addition of a speaker identifier. These preprocessing steps reduced the vocabulary to 705 user and 190 system words. For this corpus, the annotation with only the two first levels was used.

In the case of the SwitchBoard corpus, one more preprocessing step was applied. It was utterance joining: the interrupted utterances (which were labelled with '+') were joined to the correct previous utterance. No categorisation was performed because of the no-task oriented nature of the SwitchBoard corpus. After these preprocessing steps, the vocabulary consisted of 21,797 different words, which reveals that the annotation of this corpus is more difficult because of the data sparseness.

For both different annotations (Dihana two-level and SwitchBoard), the set of incremental training corpora were defined. From both training corpora, three different GIATI-based models were trained: for 2, 3 and 4-grams. These automatic labellers were used to annotate the different test dialogue corpora (Dihana two-level and SwitchBoard). The automatic annotation was compared with the reference one with the Dialogue Act Error Rate (DAER) measure. DAER (which is similar to the Word Error Rate) computes which rate of the assigned labels are correct and do not have to be revised or corrected.

Absolute results on DAER for both corpora are presented in Figure 1. As it was expected, the results are worse as the complexity of the corpus increases: Dihana is the less complex, because of the reduced vocabulary and set of labels, and SwitchBoard is the most complex (with a large vocabulary). Another clear inference from the graphics is that the larger the training set size, the better the results.

This general tendency is quite more clear with the SwitchBoard corpus, and could be related to the decrement of out of vocabulary (OOV) words as the training corpus comes larger. In Dihana the OOV reduction rate is really small for a medium-size corpus, but in SwitchBoard this reduction is higher even for a large training corpus.

One more interesting observation is that there are no significant differences between the 3-gram and 4-gram results in the Dihana corpus, but the results with 4-grams with the SwitchBoard corpus are the worst of all, while there is no significant difference between the 2-gram and 3-gram results with this corpus. The explanation is that the high complexity of the SwitchBoard corpus makes association between words and DA too sparse to appropriately infer such a complex model as a 4-gram.

In order to assess the improvement as the corpus increases, the relative improvement of passing from one training corpus to the next one in the sequence was calculated. In both corpora, these results showed that using more than 300 dialogues for training did not provide any significant improvement (lower than 5%).

Some error analysis was performed on the results with 3-grams and 300 dialogues as training corpus. The analysis revealed that most of the errors in Dihana were substitutions between similar labels or labels which annotate similar sentences but with different dialogue meaning depending on the context. This indicates that the high locality of the models do not allow to distinguish between some situations (e.g., a question and an answer). Meanwhile, in the SwitchBoard corpus most errors involved the ambiguous *statement-opinion* (sv) and *statement-non-opinion* (sd) labels, which are difficult to determine even for human annotators (Stolcke et al., 2000).

With respect to the speed of the process, the annotation technique revealed itself as really fast. In the Dihana corpus, no more than 2 seconds per whole turn on average were needed. In the case of SwitchBoard, although is quite more complex, similar times were obtained.

5 Conclusions and future work

This work shows the behaviour of an automatic dialogue annotation technique, studying the effect of the amount of training data on the accuracy of the obtained models. The experiments were carried out with very different corpora, but the results show

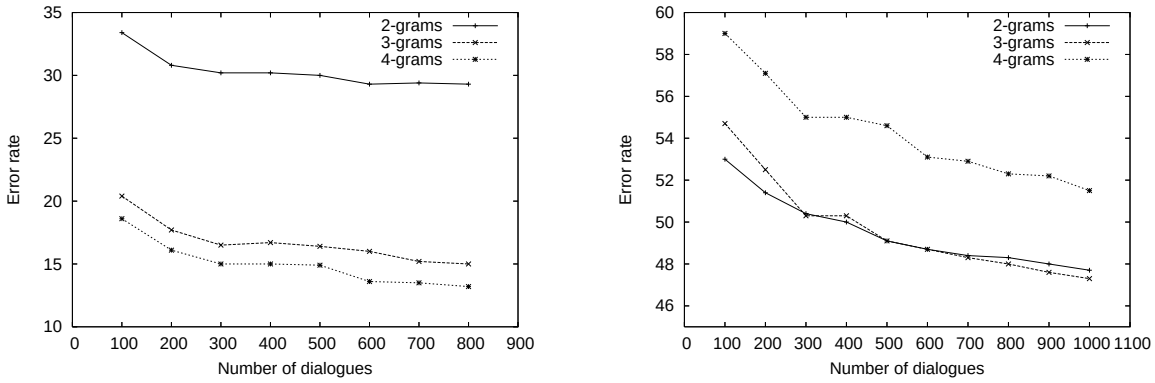

Figure 1: Absolute DAER rates for Dihana two-level and SwitchBoard.

the same behaviour: an amount of 300 dialogues is enough to obtain an appropriate annotation model. From this point, adding more dialogues to the training set does not improve significantly the accuracy of the models. Therefore, when applying this annotation technique in a dialogue corpus annotation, no new models should be inferred after the correct annotation of a relatively small number of dialogues (in this experiment, 300 dialogues). This speeds up the process, because the only task from this point is correcting the automatically annotated dialogues.

The results were obtained using the GIATI-based technique, but other annotation and identification techniques are available (Grau et al., 2004). Therefore, the same experimental framework should be applied on these techniques in order to know if they have the same limitations as the GIATI-based one. One interesting thing is the combination of several models for different tasks. Finally, although these conclusions were obtained from experiments with two corpora, experiments with more corpora could generalise these conclusions.

References

N. Alcacer, 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 SPECOM*, pages 583–586, Patras, Greece.

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

S. Grau, E. Sanchis, M. J. Castro, and D. Vilar. 2004. Dialogue act classification using a Bayesian approach. In *Proceedings of SPECOM'2004*, pages 495–499, Saint-Petersburg, Russia, September.

D. Jurafsky, E. Shriberg, and D. Biasca. 1997. Switchboard swbd-damsl shallow- discourse-function annotation coders manual. 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.

C.D. Martínez-Hinarejos. 2006. Automatic annotation of dialogues using n-grams. In *Proceedings of TSD 2006*, LNCS/LNAI 4188, pages 653–660, Brno, Czech Republic, Sep. Springer-Verlag.

J. R. Searle. 1969. *Speech acts*. Cambridge University Press.

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.