EUTRANS project: FUB activity in Spoken Machine Translation

D. Aiello, L. Cerrato, C. Delogu, A. Di Carlo
Fondazione Ugo Bordoni - Rome
{demetrio, loredana, cristina, adicarlo}@fub.it

Abstract

In this paper we report on the ongoing activity for the project ESPRIT LTR EuTrans during the first year. Issues for the Fondazione Ugo Bordoni (FUB) have been the prototype implementation, the text and speech data collection and the usability assessment. The corpus has two main objectives: to provide training material for recognition and translation training; to contain speech material for acoustic modelling, and textual material for language modelling and translation modelling. The main aim of the project is to show the potentialities of an Example Based approach to the Machine Translation for a speech-to-speech application in the domain of hotel reservation. The speech application envisaged is a telephone communication media between two human beings (the customer and the receptionist) with translation abilities: the customer call, from inside or outside, a special number of the hotel, speak in his/her own language in order to make a reservation, to cancel a reservation, to ask information for services, to claim for services; the receptionist will receive the customer request in his/her own language and will answer in his/her language through the same media. The application has been simulated through the technique of Wizard of Oz; the current prototype is based on a two-machine architecture: a recognition/translation server and client supporting the telephone interface towards the user and the GUI towards the receptionist. The WoZ plays the role of the receptionist.

Key words

Speech Machine Translation, Corpora, Example-based Method, ASR on Telephone, Visual Scenarios.

1. Introduction

Because of its great industrial and social interest, Machine Translation (MT) has received considerable attention for a long time. This paper is aimed at reporting efforts on Example-Based (EB) techniques for developing MT systems for limited-domain tasks which require text and/or speech input. We focus on the activity of Fondazione Ugo Bordoni (FUB) during the first year of the project ESPRIT LTR EUTRANS n. 30268 [1].

Prototypes relying on EB techniques for learning a kind of finite-state translation models have already been shown in a medium-sized person-to-person communication task in the hotel domain. These models lend themselves particularly well to be integrated with acoustic-phonetic, lexical and syntactic models in order to perform speech-input MT. This allows the development of systems in which all the models required for each new application are automatically learned from training data.

In order to deal with more realistic and complex translation tasks, new efforts aim currently at thoroughly exploring alternative Machine Translation approaches. We shall outline here the techniques that constitute the technological framework of the EuTrans project; namely, Finite-State Transducer Learning and Memory-Based Techniques and Statistical Translation.

On the other hand, the development of Translation Systems is heavily based on Automatic Learning techniques that extract the necessary information, structure and/or parameters from adequate corpora of training examples of each translation task considered. Clearly, this rises the problem of how to collect translation training data in such a way that it is cost-effective for limited domain applications. Specific attention is devoted to develop adequate techniques for this purpose.

2. Task Analysis and User Centred Design

With the aim of showing that proposed approaches to automatic translation are well suited for simple applications, we have defined a medium-complexity limited-domain real-world application, with spontaneous speech and a vocabulary of a few thousand words. The chosen domain is the Traveller Domain, where user's goal is to interact with a hotel reception by phone to accomplish several tasks. Task analysis helped us to choose the application domain as well as the appropriate types of user, tasks, and environment. With respect to the application domain, the Traveller Domain appeared to be reasonable for our application requirements (medium-complexity, realistic but limited-domain, spontaneous speech, a few thousand-word vocabulary). The user population of our application consists of young people who usually organise their trip without the help of a travel agent. The user's goal is to interact with a hotel reception by phone (from outside and/or inside the hotel), to accomplish five main tasks (to get information; to make a request; to cancel a request; to complain for something; to make a change) having few subtasks each. With respect to the environment, the telephone line appeared to be the most suited environment for our application.

The choice of the application domain is a very crucial point since to build effective applications it is important that they should also be useful applications, i.e. that deliver service for users. To this purpose, many suggestions should come from new approaches in the field of design and evaluation, such as the User-Centered Design (UCD) framework, the WoZ simulation technique, and the usability evaluation methods. [2]

Simulation methods such as "Wizard of Oz" (WoZ) technique are very suitable within the UCD framework. In the last few years, many laboratories and projects involved in developing spoken language systems have used the WoZ technique to elicite spontaneous spoken human-machine interaction in order (i) to collect spoken corpora and (ii) to evaluate the system in a real life condition. The basic idea behind the WoZ technique is simple: a human (usually known as the wizard or accomplice) plays the role of the computer in a simulated human-computer interaction.

The use of scenarios is crucial in WoZ simulation. A scenario is a description of an actual task that each speaker, acting as user, has to accomplish using the system. With respect to corpora collection, scenarios should stimulate speakers to produce a corpus with a large variety of words and language constructs; with respect to the evaluation of systems in real life conditions, scenarios provide subjects with realistic goals to achieve. [3]

3. Corpus collection

One of main issues for FUB contribution into the project is the training data collection. For recognition modelling, we have provided speech data with adequate phonetic transcription for acoustic modelling and orthographic transcriptions for language modelling. For the translation modelling, we have provided pairs of sentences in the source and the target languages.

In order to bootstrap the Automatic Speech Recognition (ASR) for Italian, we provided a first corpus, out of the domain, just to obtain the acoustic models. The speech material of this first corpus has been selected from a pre-existing corpus and has been adequately transcribed for acoustic modelling

In order to bootstrap the MT, text paired sentences are needed to train translation models. An initial corpus has been generated automatically starting from a set of syntactical rules.

The UCD approach suggests that either data for recognition and translation modelling have to be as natural as possible in order to realise a usable system. To this purpose, our telephonic data collection has been designed as an iterative process distributed over three periods. We have already finished a preliminary collection. The speakers were given a session guide with instructions and ten selected scenarios. They call our toll-free number to interact with our WoZ based system; all the interaction is recorded for the data collection.

By "scenario" we intend a description of an actual task that each speaker, acting as user, has to accomplish using the system. Scenarios should stimulate the subjects to produce a corpus with a large variety of words and language constructs. In general, written scenarios are used in corpus acquisition. The limit of these scenarios is that they are likely to influence a speaker in the choice of the lexical items used to express the content of the scenarios. So far, there isn't a well-known method for designing scenarios avoiding the linguistic bias introduced by written scenarios. Some work used table-scenarios, or inserted graphic representation in written scenarios. In order to give a contribution to this subject, we run an experiment to explore possible differences between written and visual scenarios. The results of this experiment showed that (i) written scenarios influence speakers in the choice of the lexicon used to express the concepts more than visual scenarios; (ii) sentences produced by speakers who received visual scenarios showed very high lexical differentiation. [4]

On the basis of these preliminary results, we have decided to use visual scenarios for our corpus collection. In order to have a high coverage of the chosen domain, we implemented a procedure that automatically produces scenario generating functions which properly describe and cover the specifications obtained with the task. [5]

3.1 Transcription of a pre-existing corpus

The first corpus provided for the project is made of 806 read passages uttered by 401 male speakers and 404 female speakers selected from the speech material a pre-existing corpus of real Italian telephonic speech collected for speaker recognition purposes.

Each read passage produced by each speaker contained is provided with an orthographic transcription. In a text file (dict.txt), which appears as a phonetic dictionary, we reported the phonetic transcription for all the lexical items contained in the production of read passages. These phonetic transcriptions follow the SAMPA conventions [6], but we deliberately decided not to transcribe the lexical accent. Conventionally we also agreed in transcribing each phoneme separated by a space:

Mario: m a r j o
cento: tS e n t o (tS is a single di-graphemic phoneme)

In order to avoid confusions between single and double consonants we decided to add the phonemes corresponding to all the double consonants of Italian in the phonological inventory (as a consequence we used an inventory with a wide number of phonemes: 48 rather than 30) and we agreed in transcribing the double consonants as one single phoneme (i.e. with no space between the repeated symbols):

scacchi: s k a kk i
accesso: a ttS e ss o

The pronunciation dictionary includes both all the words and non-words (i.e. misreading) uttered by the speakers: many mispronunciations occur; for instance for words "autorizzazione" and "bancomat", you can find non-words:

atorizzazione: a t o r i ddz a tts j o n e
autarizzazioni: a u t a r i ddz a tts j o n i
automati: a u t o m a t i
auturizzazione: a u t u r i ddz a tts j o n e
ban: b a n
bancoma: b a n k o m a
banconovat: b a n k o n o v a t

or other Italian real words coming from mispronunciations:

auto: a u t o
automatizzazione: a u t o m a t i ddz a tts j o n e
autori: a u t o r i
bacco: b a kk o
banco: b a n k o

The dictionary also reports some of the most common regional pronunciation variations realised by the speakers, which apparently could be considered as slight deviation from the canonical form, or "narrow phonetic distinctions". All the misreading (i.e. disfluencies) are reported as new entries, even if they are non-words; the regional pronunciations are reported with indexed original transcriptions, so that the transcription with index \1 always refers to the "standard" pronunciation, while all the other indexes are referred to regional or subjective variations. But there is no correlation between the number of the index and the frequency of the alternative pronunciation. The alternative pronunciation \3 is not more frequent than the alternative pronunciation \2. For example for words "autorizzazione" and "bancomat" we report the following transcriptions:

autorizzazione\1: a u t o r i ddz a tts j o n e
autorizzazione\2: a u t o r i ddz a ddz j o n e
autorizzazione\3: a u t o r i ts a ss j o n e
autorizzazione\4: a u t o r i ddz a ts j o n e
bancomat\1: b a n k o m a t
bancomat\2: b a n g o m a t
bancomat\3: p a n g o m a t
bancomat\4: p a n k o m a t

There is a particular case of regional pronunciation in which some of the alternative pronunciations are characterised by the insertion of a phoneme; this is the case of "scienze", which standard pronunciation should be /S e n ts e/, but in some of the alternative it appears the insertion of a semivowel:

scienze\1: S e n ts e
scienze\2: S j e n ts e
scienze\3: s j e n ts e
scienze\4: S j e n dz e
scienze\5: S e n dz e

When we speak of regional pronunciation we refer in particular to the phenomena of sonorisation or desonorisation of intervocalic voiceless-voiced consonants. We deliberately decided not to take into account the different distribution of open vs closed vowels.

Coarticulation phenomena at word boundaries have not been taken into account.

The dictionary counts 556 items.

3.2 New speech-input translation corpus

After the model initialization, we need more domain oriented data to optimize performances for the specific application. After the first collection session, the training corpus contains 1606 audio files for about 5h of speech produced by 181 speakers (109 males and 72 females)[1]. Speech data are stored in in µlaw format at the 8000 Hz sampling rate. Each speech file is transcribed and translated.

Some new criteria have been adopted for the transcription of the speech material gathered with the new acquisition.

We agreed that the best thing to do would be to produce a dual orthographic transcription for the two different purposes: translation and acoustic modelling.

Basically the criteria for the orthographic transcription for acoustic training remained the same used for the transcription of the previous material, with some simplification and some extension to the old convention.

It was agreed to add some "extra" information in the transcription, useful for the translation. These extra information are marked by a convention that allows to automatically take them away when not needed.

Here are the classes of new information useful for the translation:

[1] *Actually, at the date of the paper submission other data have already been collected and these numbers could be not significant of the size of the corpus.*

1) Punctuation (, ; . :!?)
2) mark of non-words that should be removed or replaced by the correct word (understanding of human-transcriber).
3) mark of false starts, hesitations and non-verbal noises that should be removed
4) Spelling sequences marked by a leading letter '$'(no cases so far).

Punctuation, non words, mistakes, hesitation to be either removed or replaced have been inserted into a "substitution pattern", which contains a string to be replaced and a replacement string delimited by slashes:

/string1/string2/

where string1 and string2 can be either null or not-null string.

We have then implemented the substitution rules in order to obtain the two transcriptions for acoustic and for translation modelling:

for translation: /string1/string2/ -> string1 (1)
for acoustic: /string1/string2/ -> string2 (2)

An example may clarify the procedure: the transcription
" mi vuol portare, per favore, la comunica la colazione in stanza? grazie."
contains some punctuation and the non sense string "la comunica" produced by mistake or by false start. The transcription with new convention will be:

mi vuol portare/,// per favore/,// //la comunica/ la colazione in stanza/?// grazie/.//

On the basis of this transcription, two transcriptions for the two different purposes can be obtained:
applying the rule (1):
mi vuol portare, per favore, la colazione in stanza? grazie.
applying the rule (2):
mi vuol portare per favore la comunica la colazione in stanza grazie

For the transcription of the pronunciation variants in this new corpus we decided to report only the most frequent variants, which apparently are mainly English words used in the Italian language and pronounced in different ways according to the speaker, for instance:
bus\1: b u s
bus\2: b a s
Also the pronunciation variants are inserted in the "substitution pattern", so the transcription in SIVA style

a che ora parte il bus\1 dell"hotel

is now with the new convention

a che ora parte il bus//\1/ dell/'/"/hotel/?//

For the production of the pronunciation dictionary for this new corpus the human transcriber can rely on the support of an automatic transcription system, that we implemented on the basis of a series of phonological rules able to perform a grapheme-to-phoneme conversion (in SAMPA conventions). The rules are entirely based on the orthographic shapes of the Italian words and they do not take in consideration any level of grammatical knowledge. [7] Although the automatic phonetic transcription is very helpful and coherent an expert phonetician has to supervise the transcriptions for exceptions and pronunciation variants.

The speech material gathered during this first acquisition phase has also been translated into Spanish and into English in order to obtain training pairs for translation modelling.

3.3 Seminal-synthetic-Text Training Data

An issue of the project is to have natural speech and text data to cope with the real input coming from the user of our "automatic translator" for the hotel reservation application. On the other hand, translation modelling, in the Finite State framework, requires much more data than whose obtained from collection with WoZ technique, then it is important to reach as soon as possible a low-performant running prototype to use it for acquisition of further data. For this purpose, in order to train a 0-level approximation of the translation model initial corpora of sentences pairs (Italian-Spanish, Italian-English) have been generated automatically.
All the subdomains included in the definition of the task were grouped mainly into two groups, topics about rooms and topics about services.

A stochastic syntax-directed translation scheme (SDTS) was developed for each one, taking Italian as input language and Spanish and English as output languages. A SDTS is a Context Free grammar in which each generation step generates simultaneously the syntactic structure for two languages. The SDTS were made providing to the input language (Italian) a wide syntactic variability based on some spontaneous speech sentences previously acquired. Against, for the output languages (English and Spanish), this variability was reduced preserving their semantic information. In other words, we have tried to reduce the output languages perplexity respect to the input one. Finally, two corpora were done, one for each pair of languages (Italian-English and Italian-Spanish).

It is worth mentioning the utilisation of word-categories, that is, pairs generated by these translation scheme include specific non-terminals representing generic numbers, hours, dates and names. Specifically we have used the following categories: NAME: whole person names, NAME_M: male names, NAME_F: female names, NAME_S: surnames, HOUR: hours, ROOM: room numbers, DATE: dates, HOTEL: hotel names, CITY: city names.

A specific SDTS was developed for each one of the categories and then it can be inserted in the main SDTS

in a very modular way. At moment these categories are quite reduced. Increasing their variability is very easy taking into account the simplicity of the SDTS associated to each category.
Besides, increasing the variability in the categories, it is a simple manner to increase the vocabulary and the complexity of the whole task.

The translation schemes described above were submitted to a generation software tool in order to generate the seminal synthetic-text training corpora: 400,000 sentence pairs for Italian-English and 400,000 sentence pairs for Italian-Spanish.

4. Collection station

The current speech-input prototype is used as a collection station; it is based on a two-machine architecture: a recognition/translation server and a telephone client. Client and server are connected on LAN via TCP/IP.

The core of the recognition/translation server is a HMM recognition system integrated with Finite State Translation in a very efficient way. The server is currently implemented on a PC Pentium running Linux Red Hat.

The telephonic client is based on commercial hardware for telephonic interface (Dialogic 41/E) in a PC Pentium under Windows NT 4.0. The client has to provide front-ends towards speakers/users and towards WoZ/Receptionist. The user front end will be a toll-free number with the reception service of a virtual hotel. It will introduce itself and then it will prompt the user for his/her request, then it will accept the request or will prompt for another formulation of the request; the interaction will be closed with some message to give the feeling of a positive conclusion of the interaction. Then the user front end will be very simple for this first prototype because it is just a sequence of prompting, acquisition and closure states. On the other hand, the front-end towards WoZ/Receptionist is constituted by a visual interface allowing the Woz to manage the call (assisted modality). The WoZ will be able to check the status of the call, to choose the most appropriate pre-recorded prompt or answer to send to the user, to edit and/or to validate the automatic recognition and translation. The client will run in automated modality too. This modality allows the collection centre to be active also in no-working hours but, of course, the interaction will be more rigid and less robust.

The speech signal on the telephone line is acquired by the telephonic board at 8000 Hz sampling rate; it is saved for collection purposes and transmitted on the network in μlaw format. Speech on the server has to be converted in an adequate format to be processed by the recognition and translation server.

The communication schema realised between client and server is very simple. A server socket process is waiting for connection on the server machine; when it accepts a connection from the client, it saves received speech data into a local file. This file is converted and then a command is sent to the server for file recognition. The server output is parsed in order to extract recognised and translated strings; another socket process is then used to send these strings to the client.

5. Conclusion

Important benefits are expected from Example-Based approaches for automatically building MT systems from training examples of each considered task. In particular it is expected to reduce the development and maintenance costs of MT systems in many specific domains but an important bottleneck is the collection of large corpora of training data. Our role in the EUTRANS project is to collect training data: about 5h of speech produced by 181 speakers have been collected, transcribed and translated to cope with recognition and translation stochastic modelling. Further collection sessions are planned and we envisage some effort in improving the size of textual material such as sentence pairs in source and target languages.

References

[1] EuTrans WEBpage: http://hermes.zeres.de/Eutrans/
[2] D.A.Norman, S.W. Draper, "User Centered System Design" Lawrence Erlbaum Associates: Hillsdale, NJ, 1986
[3] D. Aiello, L. Cerrato, C. Delogu, A. Di Carlo, M. Nisi, "Definition and Evaluation of a Speech Translation Prototype for Limited Domain Tasks", Proc. First Int. Conf. on Language Resource and Evaluation, Granada, Spain 28-30 May 1998
[4] D. Aiello, C. Delogu, R. De Mori, A. Di Carlo, M. Nisi, S. Tummeacciu, "Comparative Evaluation of Spoken Corpora Acquired by Presentation of Visual Scenarios and Textual Descriptions", 1999 IEEE ICASSP, Phoenix, Arizona, March 1999
[5] D.Aiello, C.Delogu, R. De Mori, A. Di Carlo, M. Nisi, S. Tummeacciu, , "Automatic Generation of Visual Scenarios for Spoken Corpora Acquisition", Proceedings ICSLP98, Sydney Australia, Nov. 1998
[6] Gibbon D, Moore R., Winski R.(eds.), Handbook of Standards and Resources for Spoken Language Systems, Mouton de Gruyter, Berlin - New York, 1997
[7] L. Cerrato, D. D'Alterio, A. Di Carlo, "Regole di trascrizione da grafema a fonema per l'Italiano standard", Proceedings XXVII National Workshop AIA, Genova, May 1999