Are you ready for a call? – Spontaneous conversations in tourism for speech-to-speech translation systems

Darinka Verdonik, Matej Rojc
Faculty of Electrical Engineering and Computer Science, University of Maribor
Smetanova ul. 17, 2000 Maribor, Slovenia
E-mail: darinka.verdonik@uni-mb.si, matej.rojc@uni-mb.si

Abstract

The paper represents the Turdis database of spontaneous conversations in tourist domain in Slovenian language. Database was built for use in developing speech-to-speech translation components, however it can be used also for developing dialog systems or used for linguistic researches. The idea was to record a database of telephone conversations in tourism where the naturalness of conversations is affected as little as possible while we still obtain a permission for recording from all the speakers. When recording in studio environment there can be many problems. It is especially difficult to imitate a tourist agent if a speaker does not have such experiences and therefore lacks the background knowledge that a tourist agent has. Therefore the Turdis database was recorded with professional tourist agents. The agreement with local tourist companies enabled that we recorded a tourist agent while he was at his working place in his working time answering the telephone. Callers were contacted individually and asked to use the Turdis system and make a call to selected tourist company. Technically the recording was done using PC ISDN card. Database was orthographically transcribed with Transcriber tool. At the present it includes cca. 43.000 words.

1. Introduction

The goal of speech-to-speech translation is that one day a person could for example carry out a telephone conversation with a speaker with whom she/he shares no common language. However, researchers developing speech-to-speech translation technologies (this includes speech recognition, speech centred translation and speech synthesis) find the spontaneous conversation full of phenomena like disfluencies, repairs, false starts, hesitations, filled pauses, silences, repetitions etc. In the context of the conversation, these phenomena are not disturbing, the opposite, they may have their own communicative role, but for speech-to-speech translation technologies this is very problematic. The TC-STAR consortium (http://www.tcstar.com.tw/) therefore suggests that simple combining of machine translation technics, developed for the translation of the written text, with speech recognition and speech synthesis into speech-to-speech translation systems cannot achieve satisfying quality, but special approaches to the speech centred translation are needed. Similar is concluded in the Verbmobil (http://verbmobil.dfki.de/verbmobil/VM.English.Mail.30.10.96.html) and other projects where speech-to-speech translation systems were built.

2. Recording conversations in a studio

In order to get closer to the solution, the databases of spontaneous conversations are needed first, to be able to track and study the spontaneous conversation phenomena. For that we wanted to obtain the conversations as real and natural as possible.

Recording real conversations can be difficult since speakers have to be notified in advance that their conversation will be recorded. Depending on a type of conversation we want to record, there may rise different problems. One of the problems is certainly to convince the speakers to allow the recording. We must consider that we will not get a permission to record conversations where personal or classified data are discussed. If we want to record conversations between a seller and his customer we would scare some customers since not all people are ready to be recorded, so there is a small chance that a seller would agree. We must also consider that naturalness of conversation is usually affected as soon as speaker knows that she/he is being recorded.

In many projects of speech-to-speech translation systems databases of conversations were recorded in studio: Janus (Shum et al., 1994; Lavie et al., 1997), Verbmobil (Kurematsu et al., 2000), EuTrans (Aiello et al., 1999), LC-STAR (Arranz et al., 2004) etc.), also Nespole! (Mana et al., 2004), but with professional tourist agent.

When planning the recording for the Turdis database we first recorded some examples of imitated conversations in the studio with the scenario of making hotel reservations. The speaker imitating the role of the hotel receptionist got a calendar and the speaker imitating a caller got a task to make a reservation for specified number of rooms, beds and for specified dates. The instructions were to talk to each other naturally, not to prepare sentences or learn a dialogue by heart. First recordings showed many problems. It was very hard to motivate speakers (the same report other researchers (Arranz et al., 2004). Since we were recording telephone conversations in tourist domain it was especially difficult to imitate the role of a professional tourist agent who usually provides information: if a person has never been a professional tourist agent she/he does not have all necessary background knowledge. Many problems rise from this. Some are contextual: for example a speaker in studio answered the telephone ringing with: *Dober dan. Hotel Piramida. Rezervacije.* (Eng.: *Hello. This is hotel Piramida. Reservation.*) A real receptionist at Hotel Piramida never said *reservation* in such context because they have no special desk for making reservations. Further, when comparing the conversations recorded in studio with similar conversations we later recorded through the Turdis system, we saw that conversations in studio were shorter, a speaker imitating the agent was less talkative, he provided less information to similar questions, there was no overlapping speech (there were no push-to-button restrictions), language was closer to literary standard, there were less cut-off utterances, less repetitions, repairs, silences or filled pauses, in short there were less conversation phenomena we mentioned at the beginning as problematic for speech-to-speech translation. Sometimes there appeared difficulties when speakers did not know what to say, some conversations had to be recorded more than once since speakers started to laugh or they provided wrong information and stopped, despite the instructions some speakers prepared their dialog in advance to make it more fluid etc.

3. Recording conversations with the Turdis system

In the Turdis database presented here we selected telephone conversations in tourist domain where professional tourist agent provides information to customers. The domain was chosen according to the LC-STAR project (www.lc-star.com; Arranz 2004; Foerse et al., 2004) in which University of Maribor participated as external partner and where language resources for Slovenian language were also built. The tourist domain was also the main or one of the main domains for speech-to-speech translation projects like Verbmobil, Janus, Nespole!, EuTrans...

Since the tourist domain in general is too broad as a domain of interest for typical speech-to-speech translation applications, it was further restricted to the following sub-domains:

- telephone conversations in tourist agency
- telephone conversations in tourist office
- telephone conversations in hotel reception

Comparing to the LC-STAR project we did not include telephone conversations in railway and airline companies since it would not be so interesting in the local tourist environment.

In order to avoid most of the problems rising from recording imitated conversations in studio, we made two steps: we contacted professional tourist companies for cooperation, and we enabled the speakers to use the Turdis recording system in their natural environment, professional tourist agent at his working place, and potential customer at home, office or anywhere else.

Technically this was made possible by using the ISDN card. The recording system Turdis uses both available

ISDN channels. One is used for connection with an agent and the other for connection with a caller. Callers do not call a tourist agency directly, instead they call the Turdis system. The system calls an agent in the selected tourist agency immediately after receiving a call from a caller. When both connections are established the system automatically connects both lines and establishes direct connection between a caller and an agent. At the same time recording session on both channels starts. The solution was very efficient on shorter distances, but when calling from larger distances (eg. to tourist agency 100 km or more away from where the recording system was placed) echo was already quite disturbing for the callers.

We needed two kinds of speakers for recording: professional tourist agents and callers. In order to find the agents who would participate in the recording, we first contacted local tourist companies for general permission that their agents can spend some (limited) time for recording our database at their working place during their usual working time. Then we contacted tourist agents in these companies for general permission to record the conversations they will have with the callers using the Turdis recording system. Their superiors helped us with this and there were almost no refusals.

The callers were contacted individually and asked to make a call; they were mostly employees and students of the University of Maribor. Some tries of broader campaign to find callers did not give much results, it showed out that the most successful way was to contact callers personally, one by one, if they could be motivated by promising a little reward or something in return it was even better. Callers were mostly not ready to call immediately after we asked them to, but almost everybody waited a day or two before making a call. The calls to the same tourist company were not too frequent, approximately three or four per week and in longer period, so the agents mostly did not distinguish calls through Turdis recording system from all other calls they had. This was a crucial point as we saw, since their natural reaction caused that also callers soon after starting a conversation became unaware of the recording.

We did not give much limitations about a topic of conversation since it was already enough restricted by conversational situation: calls could be made only to two hotel receptions (in Hotel Piramida and Hotel Habakuk), local tourist office (MATIC) and four different tourist agencies (Sonček, Kompas, Neckermann Reisen, Aritours), all in Slovenia. The callers were always encouraged to ask for the information they might really need or really find useful, or to think of similar calls they might had in a past. They also got the catalogues of the tourist agencies and the addresses of web pages of all the companies in case they wanted to search some information first. If callers needed help to invent an imaginary scenario for a call we prepared the list of the most common topics, collected with help of the tourist agents who participated in recordings. For tourist agencies the list of the topics was very diverse, depending mostly on a field that particular agency covers (eg. Adriatic see, pilgrimages etc.), so catalogues and web pages were very helpful. On the other hand, some topics we first thought most interesting, like making reservations, were not real for telephone conversations, since reservation in tourist agency could be done only by paying an advance. For hotel reception the most common topics were giving information about the prices, free capacities, leisure activities and congress centers; making and canceling reservations was also possible. For tourist office the most common topics were guiding tours in the Maribor, information about the city, transportation, accommodation, cultural and other events; frequently asked questions were also about driving with raft along the river Drava and ordering tourist catalogues about the city and the region.

All conversations were in Slovenian language which was also a mother tongue of all the callers. As an exemplary study we recorded one conversation between a native German speaker and Slovenian agent in the tourist office. Even though the agent spoke good German some language barriers affected the conversation, e.g. his utterances included longer pauses, sometimes searching for words, more repairing, some standard language imperfections etc.

4. Transcribing

Recorded material was transcribed using the Transcriber tool (http://www.etca.fr/CTA/gip/Projets/Transcriber/ -fr /user.html). We considered some of the EAGLES recommendations (http://www.lc.cnr.it/EAGLES96/

spokentx/) and principles of transcribing BNSI Broadcast News database (Žgank et al., 2004).

4.1 Segmentation

The basic units of segmentation of recorded material are segments or utterances. In order to achieve unitary segmentation, we defined the utterance as a semantic unit of speech, limited with pauses and marked with intonation, in speech of the same speaker. The average length of the utterances, segmented according to this rule, was between 6 and 7 words.

One or more utterances construct a turn, i.e. everything that speaker says before the next speaker starts talking. The way of recording enabled overlapping speech which is common characteristic of conversation. Overlapping speech in the Turdis database was transcribed in a special overlapping segment where speech of each speaker was transcribed. It was very common that overlapping speech started or ended in the middle of an utterance. To keep the information about utterances, we included tags that marked:

1) where the utterance is broken because overlapping speech started or ended (sign ***[1]***),

2) where the utterance continues because overlapping speech started or ended (first next sign ***[2]***),

3) where there should be a new segment for new utterance in speech of the same speaker but was not tagged by segment because of overlapping with the other speaker (sign ***[P]***).

Example (the way it is seen through Transcriber tool, not in XML file):

In Slovenian language:

Sp1[overlap]: štirinajst dni prej bi že bilo fajn

*Sp2[overlap]: mislim kak ... **[P]** štirinajst dni **[1]***

*Sp2: **[2]** najmanj prej**[+overlap_ja]** no*

In English:

Sp1[overlap]: fourteen days before is recommended

*Sp2[overlap]: I mean how ... **[P]** at least **[1]***

*Sp2: **[2]** fourteen days before**[+lex=overlap_ja]** I see*

In the recorded conversations it was very common that listener was expressing his attention, understanding, agreement with short words like *mhm, aha, ja* (Eng. *ah, oh, yeah* etc.), without signaling that he would like to take over the turn. We did not consider this as overlapping speech, so this words were tagged as special overlapping events (e.g. ***[lex=overlap_ja]*** where *lex=overlap* is the description of event and *ja* is the word that was pronounced, see previous example).

At a section level only four types of section were tagged: the beginning, the main body and the end of a conversation, and longer breaks, e.g. silences or connecting telephone line (more than 1,5 sec.).

Most of the well heard sound events were tagged: breathing, laughing, coughing, incomprehensible or quiet speech, as well as some background sounds (eg. telephone ringing).

4.2 Transcription

Transcription was basically orthographic, however phonetic transcriptions with SAMPA symbols (Zemljak et al., 2002) were added for some words. In Slovenian language there can be important difference between spoken language and literary standard because of vocal reduction. It is very common that some vowels are not pronounced when speaking, e.g. for word *[” t u: - d i]* (Eng. *also, too*) is pronounced *[” t u: t]*, instead of *[i – “ m a: m]* (Eng. *I have*) is pronounced *[” m a: m]* etc. Such pronunciations are problematic for speech recognition since the existing phonetic lexica for Slovenian language considers only literary standard pronunciation. It is usual that such pronunciations are tagged with special tag so they can be automatically tracked. In Turdis database a step further was done: for each such word a basic phonetic transcription was added in square brackets next to the word, eg. *tudi[t/u:t][+pron=*]*.

Proper names, abbreviations, spelling, foreign words... were all tagged the way they can be automatically tracked.

Some basic prosody information were added: shorter pauses within speech (sing *[.]*), prolonged syllables (sign *[:]*), raising intonation (sign *?*), emphasized pronunciation (sign #), as well as cut off words (sign *()*) and cut off or unfinished utterances (sign ...).

While Transcriber tool enables some basic information about speaker and turn, these are added as well: gender, dialect, and mode, fidelity and channel.

The format of the transcriptions in XML is determined by Transcriber tool.

5. Some main characteristics of the Turdis database

At the present the Turdis database, according to it's size, represents only a foundation for further work. The transcriptions include 43.000 words. The recorded material includes 74 recordings. In 7 recordings the agent who answered the phone connected a caller to some other agent, so there were two conversations in one recording, all together 81 conversations. The total length of the recordings is 4,6 hours, the average length of a conversation 3,4 minutes. The table 1 shows more details about number and length of conversations.

	No. of conv.	Total length
Tourist agency	48	169,2 min.
Tourist office	18	66,3 min.
Hotel reception	15	42,2 min.
Total	**81**	**277,7 min.**

Table 1: Number and total length of conversations in the Turdis database.

We can see that more than a half of the conversations were recorded with tourist agency. It is because the information that a tourist agency can offer are far more diverse and numerous than information in a hotel reception. There is a lot of tourist agencies in the local area where we were recording and many people working as tourist agents, but only one tourist office with few tourist agents providing information.

The conversations were done by 75 speakers, 42 callers and 33 tourist agents. Table 2 brings further details about the gender of the speakers. Mostly females work as a tourist agent who provides information therefore and that reflects also in the Turdis database.

	Male	Female
Tourist agents	6	18
Callers	24	18
Total	**30**	**36**

Table 2: Gender of the callers and the agents in the Turdis database.

Slovenian language is dialectically very diverse. At the most general level we distinguish the northeast dialects and the southwest dialects. The speakers in recorded conversations were speaking different northeast dialects, only three callers and three agents were speakers of southwest dialects. It is still the task for future work to obtain conversations with speakers of the southwest dialects.

6. Conclusion

The Turdis database was basically built for use in developing speech-to-speech translation components: speech recognition, speech centred translation, speech synthesis. At the moment it can be used for improving acoustic and language models. It is very important source of knowledge about conversation and spontaneous speech phenomena (eg. vocal reduction, morphological particularities, conversational words which are not part of literary standard, utterance structure like repairing, non-standard word order) which is necessary for developing technology. It can be used also for developing dialog systems or for linguistic researches. The next step should be translation of the transcriptions in Turdis; as the parallel corpus it could be used also for developing speech centred translation.

7. Acknowledgements

We sincerely thank to all tourist companies who participated in recording: tourist agencies **Sonček, Kompas, Neckermann Reisen** and **Aritours**, to **Terme Maribor**, especially **Hotel Piramida** and **Hotel Habakuk,** and to **Mariborski zavod za turizem** with tourist office **MATIC**. We also thank to all tourist agents in these companies who participated in recording and to all the callers who were ready to use the Turdis system.

8. References

Aiello, D., L. Cerrato, C. Delogu, A. Di Carlo (1999). The Acquisition of a Speech Corpus for Limited Domain Translation. In Proceedings of Eurospeech 1999, Budapest.

Arranz, V., N. Castell, J. Gimenez, H. Ney, N. Ueffing, (2004). Description of language resources used for experiments. http://www.lc-star.com/archive.htm.

Foerse, H., E. Hartikainen, H. van den Heuvel, G. Maltese, A. Moreno, S. Shammass, U. Ziegenhain (2004). Creation and Validation of Language Resources for Speech-to-Speech Translation Purposes. In Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal.

Kurematsu, A., Akegami, Y., Burger, S., Jekat, S., Lause, B., MacLaren, V., Oppermann, D., Schultz, T. (2000). Verbmobil Dialogues: Multifaced Analysis. In Proceedings of the International Conference of Spoken Language Processing.

Lavie, A., L. Levin, P. Zhan, M. Taboada, D. Gates, M. Lapata, C. Clark, M. Broadhead, A. Waibel (1997). Expanding the Domain of a Multi-lingual Speech-to-Speech Translation System. In Proceedings of the Workshop on Spoken Language Translation. Karlsruhe, Germany.

Mana, N., R. Cattoni, E. Pianta, F. Rossi, F. Pianesi, S. Burger (2004). The Italian NESPOLE! Corpus: a Multilingual Database with Interlingua Annotation in Tourism and Medical Domains. In Proceedings of 4th International Conference LREC'04. Lisbon, Portugal.

Shum, B., L. Levin, N. Coccaro, J. Carbonell, K. Horiguachi, R. Isotani, A. Lavie, L. Mayfield, C. P. Rose, C. Van Ess-Dykema, A. Waibel (1994). Speech-Language Integration in a Multi-lingual Translation System. In Proceedings of AAAI Workshop on Integration of Natural Language and Speech Processing.

Zemljak, M., Kačič, Z., Dobrišek, S., Gros, J., Weiss, P. (2002). Računalniški simbolni fonetični zapis slovenskega govora. Slavistična revija, 50(2), 159--169.

Žgank, A., T. Rotovnik, M. Sepesy Maučec, D. Verdonik, J. Kitak, D. Vlaj, V. Hozjan, Z. Kačič, B. Horvat. Acquisition and Annotation of Slovenian Broadcast News Database (2004). In Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal.