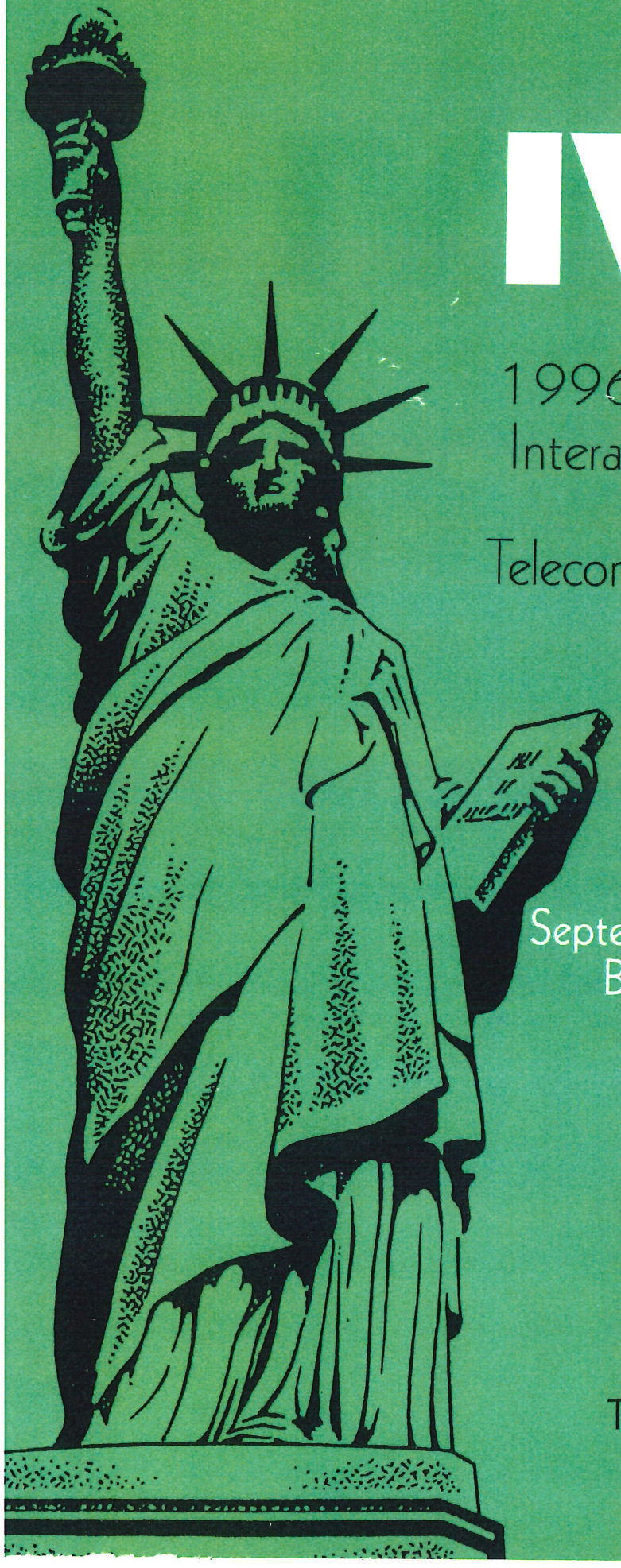

IVTTA

1996 IEEE Third Workshop Interactive Voice Technology for Telecommunications Applications

September 30 - October 1, 1996
Basking Ridge, New Jersey

IEEE COMMUNICATIONS SOCIETY

Sponsored by
The IEEE Communication Society

96TH8178

Approaches to Gathering Realistic Training Data for Speech Translation Systems

Ivan Bretan, Robert Eklund and Catriona MacDermid
Telia Research AB, S–136 80 Haninge, SWEDEN

Abstract — The Spoken Language Translator (SLT) is a multi-lingual speech-to-speech translation prototype supporting English, Swedish and French within the air traffic information system (ATIS) domain. The design of SLT is characterized by a strongly corpus-driven approach, which accentuates the need for cost-efficient collection procedures to obtain training data. This paper discusses various approaches to the data collection issue pursued within a speech translation framework. Original American English speech and language data have been collected using traditional Wizard-of-Oz (WOZ) techniques, a relatively costly procedure yielding high-quality results. The resulting corpus has been translated textually into Swedish by a large number of native speakers (427) and used as prompts for training the target language speech model. This "budget" collection method is compared to the accepted method, i.e., gathering data by means of a full-blown WOZ simulation. The results indicate that although translation in this case proved economical and produced considerable data, the method is not sensitive to certain features typical of spoken language, for which WOZ is superior.

INTRODUCTION

The Spoken Language Translator (SLT) is a speech-to-speech translation prototype capable of translating utterances from the domain of an air traffic information system (ATIS). Currently, the system translates spoken English into spoken Swedish and French as well as spoken Swedish into English and French speech, using a vocabulary of approximately 1200 stem entries [1].

SLT was conceived using an existing speech recognition system, DECIPHER as the point of departure [2]. DECIPHER has been trained for the ATIS domain using data collected in a large-scale Wizard-of-Oz (WOZ) simulation by the MADCOW (Multi-site Atis Data COllection Working) group [3] to be described below. These data are needed both to train the acoustic-phonetic model of the speech recognizer and the *n*-gram language model. In addition, they can be used to stream-line the development of the linguistic modules of the system, in particular the lexicon, grammar, set of collocations, transfer rules, and dialogue model (if existing). This can be achieved rationally by means of constructing representative corpora, where utterances are sorted according to the frequency of their syntactic pattern [4].

In a corpus-oriented development framework, the quality of the system is dependent on the quality of the corpora used. Thus, in SLT, we have devoted significant efforts to obtaining high-quality linguistic data. A question which must be answered before embarking on such undertakings is what the measures of quality are. One could imagine that genuine human–human conversations would provide the best yardstick for linguistic training data, but this is not true given that the linguistic performance models for users engaging in dialogue with a machine varies with the behaviour of the system [5]. There is also the ethical issue of "bugging" people's conversations, which is most critical if the speakers could be identified and if the dialogue contains sensitive or personal information. Ideally, informed consent should then be obtained. These facts explain why WOZ simulations continue to constitute the preferred means of collecting linguistic training data. However, WOZ simulations have drawbacks—most notably they are laborious, time-consuming and costly. An informal figure quoted in the work on collecting ATIS data estimates the cost of collecting 10000 sentences to USD 1 million, i.e., $100 per sentence! Although the WOZ simulations we have conducted ourselves for Swedish indicate much lower costs, it is clear that this way of collecting data is expensive.

The question then arises whether it is worth initiating a new WOZ simulation for each new language added to SLT. As a cheap substitute approach we have instead been experimenting with "piggy-backing" on the existing American data through textual translation by native Swedish speakers. The resulting data served as the language model for the Swedish version of SLT, by having speakers read utterances from the text corpus for the training phase.

WIZARD-OF-OZ SIMULATIONS

In order to collect more realistic training data for a spoken dialogue system, experimental subjects can be recorded as they conduct task-oriented dialogues with a simulated dialogue system. The subjects are most often led to believe that their dialogue partner is a prototype system, when in fact an accomplice (the 'wizard') is simulating an operational system by performing some or all of the system's functions. Typically, the wizard interprets the subjects' utterances, simulating speech recognition and often language understanding. Other functions, such as dialogue management and speech synthesis, can be handled by a computerized tool operated by the accomplice [6].

0-7803-3238-5/96$5.00©1996

Since the subjects have no real task they wish to complete, they are given scenarios describing a given task. Written scenarios are most common, although these can act as a 'script', strongly influencing the subject's vocabulary and syntactical structures, at least in their opening utterance within the dialogue. To overcome these limitations, the scenario can be presented in tabular or graphical form, such that the subject has to interpret the scenario using their own words. In this case, the illustrations must be unambiguous for the subjects.

AMERICAN ATIS SIMULATIONS

One example of a WOZ simulation (but with text instead of spoken output from the simulated system) is MADCOW. A large number of subjects (2724) were given written scenarios and spoke to what they believed to be a working system, when in fact human wizards were interpreting the subjects' questions, querying the database by hand, and displaying the results on the subjects' screen. After several months, once enough data was acquired in this way to train the system, the system became fully operational, and the wizards were no longer needed. The expectation was that, because the users believed that they were speaking to a real system, the wizard data and real data were equivalent.

SWEDISH ATIS SIMULATIONS

A spoken translation system in the ATIS domain has recently been simulated at Telia Research in Sweden without the use of any computerized simulation tool. Subjects believed that the 'system' translated their Swedish enquiries to an English-, French- or German-speaking travel agent in order to book flights. In fact, no translation occurred in these dialogues. The subject's utterance was conveyed—usually verbatim—by a wizard (W1), a professional actor, representing the subject's translation system, to a second wizard (W2) representing the travel agent's translation system. Certain simplifications were made to complex utterances, that is, utterances that were not understood or that were longer than twenty words were rejected by W1, who asked the subject to repeat or reformulate the utterance. The utterance was then conveyed by W2 over the telephone to a Swedish accomplice (the 'travel agent'), who asked for additional parameters where necessary to complete the booking. W2 then conveyed the travel agent's replies to the subject via W1 according to the same constraints. The two wizards sat in the same room and when they spoke to each other to convey utterances or clarify internal misunderstandings, W1 suspended the microphone contact with the subject. Similarly, W2 used the secrecy button on the telephone. The wizards listened to the respective dialogue partners through headsets.

The actor was trained to speak to the subject with the unnatural prosody characteristic of digitized speech and had no script apart from the requests to reformulate or repeat. The 'travel agent' used certain standard phrases based on an interview with a real travel agent and had access to a paper database constructed with data from an ATIS-type database used in the travel agency. Otherwise, his speech was spontaneous in response to the subject's queries, which were based on a combination of written and graphical scenarios. The dialogue between the subject and the actor was recorded as sound files on a Unix work station and all four input channels were recorded on a DAT-recorder for later transcription (see Fig. 1).

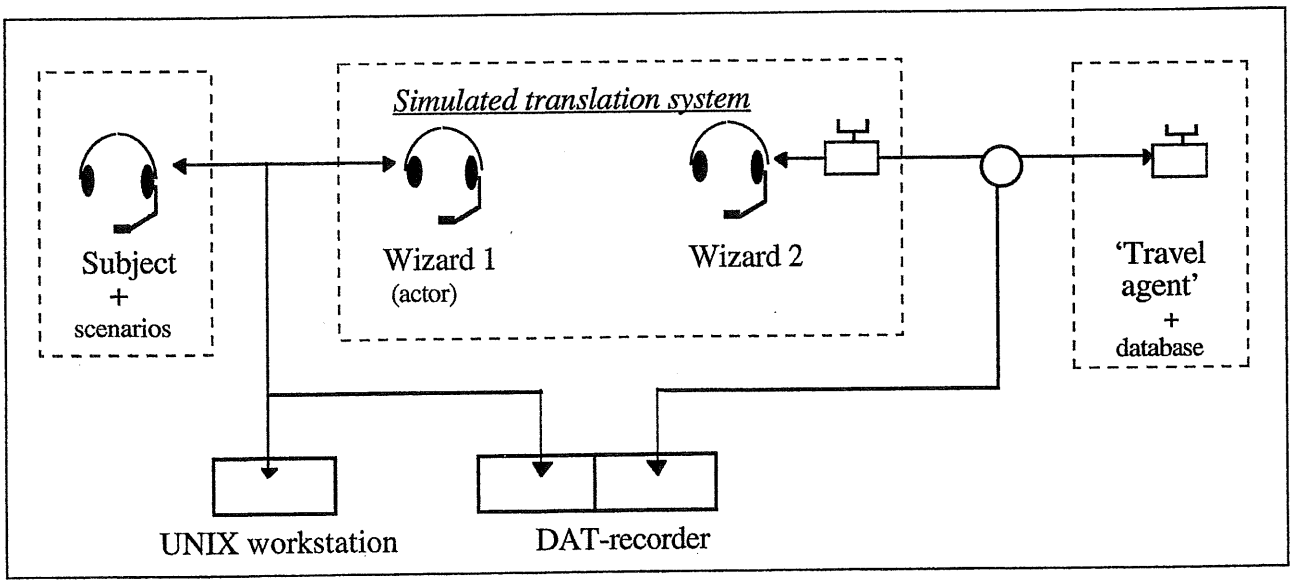

Fig. 1. Experimental set-up for Swedish ATIS simulation

TRANSLATIONS OF AMERICAN WOZ MATERIAL

One way to obtain data without having to design WOZ simulations is to translate existing WOZ data collected in another language. This was also done at an early stage in the SLT project, where 4021 sentences from the American ATIS corpus were translated into Swedish by a bilingual secretary. (This corpus will be referred to as C1). Another 3578 sentences were added by a member of the SLT team (C2), who was told to use free translations.

However, there are several problems associated with this method. First, it is very hard for translators to avoid linguistic bias caused by the wordings in the source language, and translations are almost certain to be influenced by expressions idiomatic in the source language but not in the target language. Second, phenomena like hesitations, repairs, false starts and so forth are not readily translated in a natural way. Moreover, translation is likely to miss certain features typical of spontaneous speech, like agreement errors, especially if the target language does not make use of agreement in the same way as the source language. If the target language has different grammatical granularity than the source language, translators will not add grammatical errors that do not exist in the source language. Third, one sole translator is not likely to be able to provide as great linguistic variability as will the would-be users of the system in question.

To solve these problems, it was decided to use as many translators as possible. To this end, lists of (Telia) email addresses were obtained. In a first batch, 11232 sentences from the American Atis corpus, were divided into files of 18 sentences each. These files were then e-mailed to 624 persons with accompanying instructions for translation as well as some background information concerning the task.

The translators were instructed to read each English sentence, consider its meaning, fill in the corresponding Swedish entry line and return the mail by using "Reply" in their mail client. The translators were instructed to aim for idiomatic rather than literal translations, i.e., wordings that the translators would use themselves. They were also informed that the addressee of the utterances would be a computer and not a human being. Thus, although subjects were instructed to use natural speech, they were to avoid excessive slang. In this batch, 1116 sentences were returned with translations, i.e., roughly 10 % of the mailed material.

In order to improve that figure and to obtain more data, a second batch of 7533 sentences was mailed to 2511 persons, each person receiving only three sentences each, i.e., a much lesser undertaking. This time, 360 files were returned with translations, i.e., around 14.5 %. Five additional persons translated between 18 and 100 sentences each, thus augmenting the corpus with some 500 sentences. The resulting "email corpus", after editing, contained 4595 sentences translated by approximately 427 different persons.

A COMPARISON OF THE CORPORA

The different corpora thus collected may vary according to several parameters, such as lexical size, grammatical coverage and idiomaticity, i.e., the use of idiomatic expressions specific to the domain and language.

In all the comparisons, C1 and C2 were merged into one corpus, TC (two translators), the email corpus will be referred to as EC (427 translators). A small Wizard-of-Oz pilot of 127 sentences will be called WOZp (10 subjects), whereas the WOZ simulation described above will be called WOZ (52 subjects).

One issue to be examined here is lexical representation. TC contains 1573 lexical entries (types). Here, inflected forms etc. are counted as different types. EC contains 1789 entries, WOZ contains 977 entries and WOZp 174 entries. Fig. 2 shows how the lexicon grows as a function of the number of collected sentences.

As is seen, the lexicon grows most rapidly in EC, whereas the growth rate is more or less the same for TC and WOZ. This seems to indicate that a fast way to obtain good lexical coverage is to involve many persons in the gathering of data for the target language.

Another consequence of this is probably reflected in the percentage of words in the lexicon that occur only once in the corpus. In TC and WOZ, around 10 % of the lexical entries have only one token, whereas the corresponding figure for EC is 17 %.

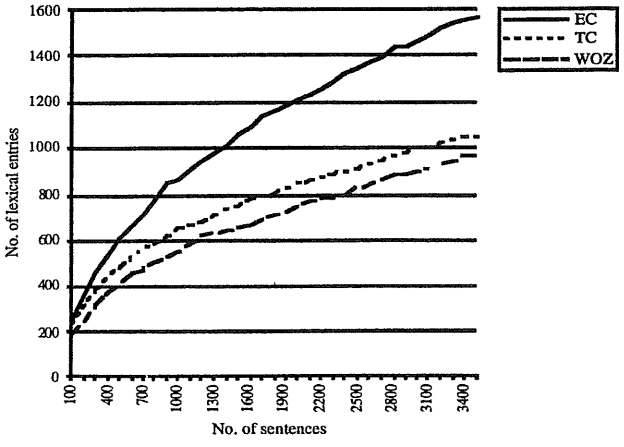

Fig. 2: Lexicon growth as a function of number of sentences. To avoid sequential effects the number-of-sentences data were collected in different orders that were averaged.

It does not follow from the fact that EC has a larger number of lexical entries than WOZ that EC is the most representative of Swedish usage, due to the aforementioned problems of 'colouring effects' and the lack of speech-specific phenomena associated with textual translation. In fact, the two lexica vary in several respects. First, there are words and constructions that exist in both lexica, but whose frequency is quite different. As an example, the word "okay" exists in both lexica, but is far more common in the WOZ material (sic!). It exists in 3.8 % of the sentences in WOZ and 2.5 % in WOZp, but only in 0.2 % of the sentences in EC. The opposite is equally true. The word "vänlig" (literally "friendly") is far more common in EC (4.8 %) than in WOZ (0.2 %), probably as an effect of "please", a word lacking a good counterpart in Swedish. What Swedes would say is "tack" ("thank you") in sentence-final position. Thus, "tack" occurs 20 % of the sentences in WOZ but only in 2 % of the sentences in EC. The corresponding figure for WOZp is 11.5 %. Since these words have different syntagmatic properties, they also influence grammatical structure.

Second, some words and/or constructions common in WOZ do not exist at all in EC. A striking example is that 7 % of the sentences in WOZ begin with the word "då" or "ja, då" (fillers that roughly translates as "well"), but not at all in EC. The corresponding figure for WOZp is 0.6 %. The idiomatic expression "det går bra" (literally "it goes well", i.e., "that's fine") occurs in both WOZ and WOZp (c. 0.7 %) but not in EC.

Similar examples of skewed material abound, most of which can be accounted for in terms of linguistic bias associated with the translation process. However, it must be borne in mind that the different set-ups for the American ATIS simulations and the Swedish WOZ simulations beyond doubt influenced the linguistic material obtained.

CONCLUDING REMARKS

One problem with the e-mail approach is that disfluent or 'strange' sentences are less likely to be translated than 'normal' sentences, since the former require more effort from the translator. This means that the method might act as a filter where marginal sentences become underrepresented in the translation process.

The costs of the method are hard to judge, since the work is very much "hidden". More than 400 persons worked approximately 10-30 minutes each, on a voluntary basis at no cost to the project. The bulk of the work consisted of editing the returned files, many of which did not arrive in the desired format.

Although the e-mail approach produced useful data and translators were instructed to respect source language disfluencies and spoken language style, the results differed from the data obtained in the Swedish WOZ simulation in that certain features of natural speech were notable by their absence in the former corpus. These differences can be attributed partly to a loss of naturalness in the translation process but more importantly to the fact that typical spoken language phenomena [7] are very specific to language and modality and cannot be obtained through literal translation of text.

One way to circumvent this problem is to record oral translations from text. A method where this is used is the Storyboard method. In this method, subjects are given picture or 'storyboard' scenarios and asked to formulate an equivalent utterance [8]. Linguistic bias from written scenarios is avoided. The data is gathered as speech rather than in written form, adding realism, though subjects are not in a 'live' dialogue. Consequently, the method is best suited to collecting initial utterances and is a good way to tap possible variations in use of syntax and vocabulary.

In conclusion, our recommendation would be to use WOZ simulations to obtain natural speech data, complemented by the e-mail approach—or a similar method—where the task is distributed among a large number of people proficient in both languages to obtain wide lexical coverage.

ACKNOWLEDGMENTS

We wish to thank Camilla Eklund, Anna-Lena Ereback, Kate Hunicke-Smith, Jaan Kaja, Inge Karlsson, Martin Keegan and Manny Rayner and hundreds of subjects and translators.

REFERENCES

[1] M.-S. Agnäs et al., "Spoken Language Translator: First-Year Report", *SICS Research Report R94:03*, Swedish Institute of Computer Science, Stockholm, Sweden, 1994.

[2] H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, "Large Vocabulary Dictation Using SRI's DECIPHER (TM) Speech Recognition System: Progressive Search Techniques", *Proc. Int. Conf. on Acoust., Speech and Signal*, Minneapolis, Minnesota, 1993.

[3] C. T. Hemphill, J. J. Godfrey and G. R. Doddington, "The ATIS Spoken Language Systems Pilot Corpus", *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, 1990. `ftp://jaguar.ncsl.nist.gov/madcow/logs/atis-spkr-info.log`

[4] M. Rayner, P. Bouillon and D. Carter, "Using Corpora to Develop Limited-Domain Speech Translation Systems", *Proc. Translating and the Computer 17 (ASLIB)*, November 1995.

[5] I. Bretan, A.-L. Ereback, C. MacDermid, C. and A. Waern, "Simulation-based Dialogue Design for Speech-Controlled Telephone Services", Proceedings CHI'95, Denver, April 1995.

[6] R. Amalberti, N. Carbonell and P. Falzon, "User representations of computer systems in human-computer speech interaction", *Int. J Man-Machine Studies* vol. 38, pp. 547-566, 1993.

[7] V. Fromkin (Ed.) *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand*, New York Academic Press, 1980.

[8] C. MacDermid and M. Goldstein, "The 'Storyboard' Method: Establishing an unbiased vocabulary for keyword and voice command applications", *Proc. HCI'96*, London, August 1996, in press.