MULTILINGUAL PERSON TO PERSON COMMUNICATION AT IRST

B. Angelini, M. Cettolo, A. Corazza, D. Falavigna, G. Lazzari
IRST - Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo, Trento, Italy

ABSTRACT

This paper refers to a machine-mediated person-to-person multilingual communication system. Stress is put on *robustness*, that is the ability of the system to preserve communication even in presence of the variability and errors typical of spoken language systems. The statistical approach is adopted not only at the acoustic level, but also for the linguistic processing. Therefore, while an overview of the global architecture will be briefly introduced, the focus will be put on the acoustic recognizer and the understanding module. Experimental evaluations complete the presentation.

1. INTRODUCTION

This paper refers to a machine-mediated person-to-person multilingual communication system. The scenario involves appointment negotiation between two persons speaking different languages. The focus will be put on two of the system modules, while an overview of the global architecture will be briefly introduced.

The expression *multilingual communication* is preferred to speech-to-speech translation in this context. In fact, the goal is to only transfer the semantic contents necessary for the communication rather than a detailed translation of the input utterance. Therefore, a lot of problems regarding the construction of a detailed linguistic representation necessary for translation are not addressed. On the other hand, stress is put on *robustness*, that is the ability of the system to preserve communication even in presence of the variability and errors typical of spoken language systems.

An important feature of all speech-based interfaces is the ability to cope with spontaneous speech at both acoustic and linguistic levels. Since in our system all modules are based on a statistical approach, it is of great importance to collect a corpus including a meaningful set of such spontaneous phenomena. In this first phase, person-to-person conversations in the same language have been recorded, in order to collect speech data as natural as possible.

As long as robustness is concerned, the two most critical modules are the acoustic recognizer and the understanding module.

At the acoustic level, our approach has been to start with the recognizer which has been developed for dictation tasks and to adapt it to the new task. The work is still in progress, and only the few first steps are here evaluated, while the following directions will be proposed in the last section.

The core of the *understanding module* is a statistical algorithm, as will be discussed. The adoption of a statistical approach allows the integration of information about dialogue history in the understanding process. Experimental results will show obtained improvements.

The following section gives an overview of the system architecture. Then, Section 3 briefly describes the data collected so far, which have been used in the experiments described in the rest of the paper. The successive sections describe the two modules above mentioned: Section 4 discusses the strategies adopted by the acoustic recognizer, while Section 5 deals with the semantic classifier. Some final remarks on the results obtained so far and future research direction complete the paper.

2. SYSTEM OVERVIEW

The architecture of the system is depicted in Figure 5. It communicates with analogous systems developed for other languages by using a formal representation of the input meaning, as much as possible language independent (Interchange Format, IF). As long as only the contents necessary for communication are transferred, a very simple IF will be used, mainly represented by a collection of Dialogue Acts (DAs) and some parameters. In Section 5, the strategy adopted to associate a DA to every input segment is discussed.

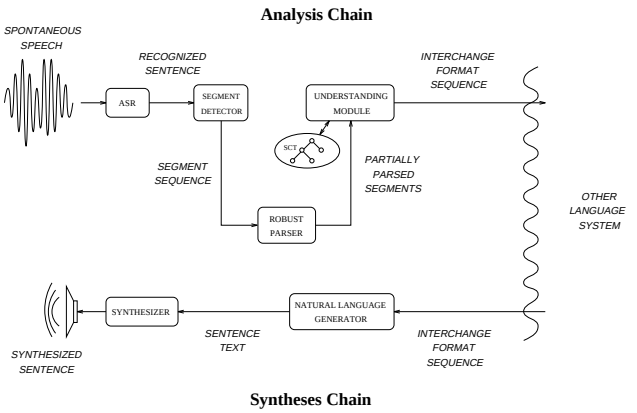

Figure 1. System architecture.

Two main processing chains can be devised: one for analysis and the other for synthesis. The analysis chain converts the input speech signal into its IF representation, through the segmentation of the recognizer output into a sequence of segments, each of which corresponds to a DA. The synthesis chain produces a synthesized audio message expressing the input IF instance.

3. CORPUS DESCRIPTION

Since IRST is only concerned with the parts of the system connected with Italian language, the data were collected by monolingual person-to-person conversations. The two speakers were asked to fix an appointment, observing the restrictions shown on two calendar pages they were given. The conversations were naturally uttered by the speakers, without any machine mediation. It is our opinion that the introduction of a machine-based interface will influence the speaking attitude of the users, simplifying the language and controlling the spontaneous speech phenomena. Therefore, the data collected in the first phase will be used to build a start-up system, to be used for further simulations. Such simulations will then be performed in more realistic conditions, and the collected data will increasingly approximate on-the-field conditions.

Dialogues took place in an acoustically isolated room. The two subjects did not see each other and could hear the partner only through headphones. Each of the two speech signals was acquired by two microphones, one directional and one close-talk, and recorded on a digital audio tape.

The speakers participating in the experiment were 27: 14 females and 13 males. They were divided in nine groups of three: for each group, 6 dialogues were recorded, one for every ordered pair of speakers, for a total of 54 dialogues.

The dialogues were transcribed by annotating any extra-linguistic phenomena like mispronunciations, restarts and human noises, but not pauses. The transcription allowed calculating the figures in Table 1 that characterize the corpus.

The average number of turns per dialogue is similar to that declared for analogous tasks [8, 7], while the total number of words in the corpus (|W|) results in a smaller average length of turns. The perplexity (PP) was estimated on a smoothed bigram Language Model (LM) with no word classes.

Table 1 also reports some figures on the most frequent extra-linguistic phenomena occurring in the corpus.[1] Only 210 sentences (21.1%) are not affected by any such phenomena.

Furthermore, 75 dialogues (601 sentences) collected at CMU[2] in American English were translated into Italian and used for preprocessor development.

# dialogue	54	
# turn	997	18.5 turn/dialogue
|W|	10586	10.6 word/turn
|V|	915	
PP	77.3	$(Pr[62.8, 95.9] = 0.95)$
FREQUENT HUMAN NOISES		
phenomenon	# occurr.	# affected sentences
inhalation	406	298 (29.9%)
vowel lengthening inside words	320	240 (24.1%)
exhalation	295	236 (23.7%)
eee	262	198 (19.9%)
mouth	218	184 (18.5%)
mmm	80	74 (7.4%)
eh	58	56 (5.6%)
ah	35	34 (3.4%)

Table 1. Corpus statistics.

4. ACOUSTIC RECOGNIZER

The recognizer is derived from that developed at IRST for large vocabulary dictation tasks (20K words). It is based on 48 context and speaker independent phonetic units that are modeled by left-to-right HMMs, with three or four states, trained on the phonetically rich database, APASCI, collected at IRST [1]. The recognizer provides the best sentence as output.

On the 1K-word vocabulary, a Shift-β bigram LM was estimated [6]. Intra-words acoustic constraints (phonetic transcriptions) and inter-words linguistic constraints (language model) are compiled into a sharing-tail tree-based network that defines the search space, at unit level, for the decoding algorithm [3].

Since little data are available for LM training, eight word classes were defined: *forename*, *place-name*, *day-name*, *meal-description*, *month-name*, *-ten2nineteen-* (numbers between "ten" and "nineteen"), *-tens-* ("twenty", "thirty", ..., "ninety"), *-units-* ("one", ..., "nine").

A preliminary recognition experiment [2] was performed on the whole collected corpus. Given a test sentence, the LM was estimated on the remaining 996 in addition to the 601 translated sentences. No model of extra-linguistic phenomena was used. Word error rate (WER) 45.1% and sentence error rate (SER) 84.1% were obtained.

Afterwards, to ease in experiments, a test set was selected, such that recognizer performance for it was very similar to the global recognition rate. It is composed by all and only the sentences uttered by 4 speakers (1 female and 3 males), for a total of 115 sentences. The rest of the corpus (transcriptions) was then used to train the bigram LM. In the following, if not otherwise specified, the term *test set* will always refer to these 115 sentences corresponding to 4 speakers, while the term *training set* will designate the rest of the corpus.

One of the problems that need to be faced when adapting a recognizer developed for dictation to a spontaneous speech corpus is the adaptation of acoustic units to the new task. Then, the original models, only trained on the APASCI database, were adapted to the new training set.

Units	WER	SER
Apasci	44.3	85.2
Means adaptation	39.2	83.5
Model re-training	36.3	81.7

Table 2. Word and sentence error rate of the recognizer using different units.

A first experiment was performed, in which the means of Gaussian probability densities of the HMMs were adjusted using the training corpus, according to the following relation:

$$\mu_{adapted} = \frac{c}{\tau + c}\mu_{NEW} + \frac{\tau}{\tau + c}\mu_{AP}$$

where μ_{AP} is the mean trained on the APASCI database, μ_{NEW} is the corresponding mean estimated on the new corpus, c is the number of samples used during re-estimation, and τ is a heuristic coefficient.

Afterwords, a complete re-training on the new data was performed, where the APASCI models were only used as bootstrap. In both cases, all the extra-linguistic phenomena were collapsed in a unique acoustic unit, which in the

[1] Note that pauses are not considered.

[2] The authors would like to thank the JANUS group at CMU for giving us these dialogue transcriptions.

APASCI data was only used for silence. Results are presented in Table 2.

5. SEMANTIC CLASSIFICATION

The core of the understanding module is represented by a classifier, based on Semantic Classification Trees (SCTs) [4, 5], which associates a DA with each input segment. In addition to the most likely DA, it also gives a probability distribution on the DAs set, to be used if DA hypotheses are to be scored on the basis of different information sources (acoustic, syntactical, and so on). Eventually, the most likely DA will be chosen. At the moment, SCT probability distribution is integrated with the probability of the dialogue history, i.e. the last DAs that occurred.

Data are preprocessed in such a way that some words and phrases are clustered and replaced by a label. A special preprocessor was used to detect temporal expressions.[3] In addition to that, the preprocessor clusters some typical expressions and sentence variations that are in fact equivalent.

The classification is performed on the basis of keywords automatically extracted from the training corpus by the tree construction algorithm. By collapsing the different expressions of the same phenomena, the preprocessing copes with data sparseness and makes the statistics used by SCTs more reliable.

In the experiments, a corpus of 1383 segments is used, resulting from the manual semantic segmentation of the collected data. Such segments have been labeled using the 18 DA labels shown in Table 3 together with the number of their occurrences in the corpus and the corresponding frequencies. In the tests, the Leaving-One-Out (LOO) technique was used for each dialogue: the segments of the current dialogue were classified using a tree grown on the other dialogue's segments. Classification errors are 535, corresponding to a 38.7% error rate.

DA labels	# occurr.	distribution
request-response	265	19.2 %
state-constraint	166	12.0 %
suggest	142	10.3 %
affirm	141	10.2 %
closing	114	8.2 %
confirm-appointment	72	5.2 %
task-definition	71	5.1 %
opening	66	4.8 %
acknowledge	57	4.1 %
accept	54	3.9 %
reject	50	3.6 %
garbage	45	3.3 %
request-suggestion	45	3.3 %
negate	33	2.4 %
request-clarification	24	1.7 %
information	15	1.1 %
clarification	12	0.9 %
request-information	11	0.8 %
total	1383	100.0 %

Table 3. Distribution of DAs in the corpus.

In a second experiment, only seven DAs were considered, obtained by clustering more similar classes. In Table 4 cluster definitions and their occurrences in the corpus are reported. Adopting again the LOO technique, 364 classification errors were obtained, resulting in a 26.3% error rate. Table 5 reports the corresponding confusion table.

cluster	DAs	# occurr.
agree	accept affirm	195
give	clarification information	27
greet	opening closing	180
propose	request_response suggest confirm_appointment task_definition	550
refuse	state_constraint reject negate	249
require	request_suggestion request_clarification request_information	80
waste	garbage acknowledge	102
total		1383

Table 4. DA clusters and their occurrences.

hypot. cluster	*true cluster*						
	agr	giv	greet	prop	ref	req	was
agree			2	44	1	6	2
give	1			22	3		1
greet	4			1		7	5
propose	27	4	7		22	11	20
refuse	1		1	32		2	6
require	4		1	26	1		18
waste	11	1	2	28	16	24	
err (#)	48	5	13	153	43	50	52
err (%)	24.6	18.5	7.2	27.8	17.3	62.5	51.0

Table 5. Confusion matrix of DA clusters.

The SCT classifier is only part of the understanding module, that, as described in [4], integrates this statistical algorithm with knowledge-based mechanisms. This is particularly effective in solving rare situations in which statistics are weak.

5.1. Dialogue History Integration

The information taken into account by the SCTs only includes the text of the sentence. One important information source which could effectively help classification is represented by the dialogue history. Indeed, in a person-to-person dialogue, there is a strong dependence between the current DA and what both users have told until that moment. Since this information is available, it can be integrated with the classifier output.

Figure 2 depicts the adopted strategy, which is based on two different conditional probability distributions defined on the DA set: the first one, $\Pr(DA \mid \text{history})$ is identified by the sequence of previous DAs; the second one, $\Pr(DA \mid \text{words})$, only by the current word sequence. The former is the probability distribution given by trigrams of DAs; the latter was estimated by using SCTs.

Three different strategies were tested to integrate the two probabilities. All of them assume that the two distributions are independent, which seems reasonable in the case considered.

[3] The authors would like to thank Alberto Lavelli for such temporal expression detector.

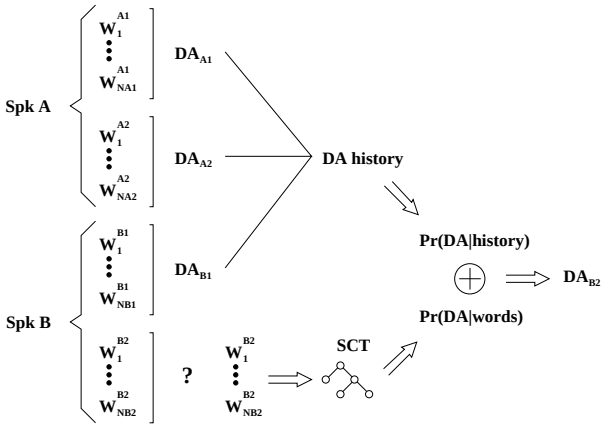

Figure 2. Scheme for dialogue history integration into the SCT-based segment classification.

classifier	error	ΔE (wrt *)
$\Pr(DA \mid \text{words})$	**38.7 %**	-
$\Pr(DA \mid \text{history})$	63.3 %	-
$\Pr(DA \mid \text{words})^{\lambda} \cdot$		
$\cdot \Pr(DA \mid \text{history})^{1-\lambda}$	35.1 %	-9.2 %
N-best rescoring	37.2 %	-3.9 %
δ-best rescoring	34.9 %	-9.9 %

Table 6. DA classification results by using single KSs and different types of integration.

In the first strategy, a linear combination of the two log-probabilities is used to evaluate the probability of each DA. The DA having highest probability is then output. In the other two, a two-step approach is adopted: a restricted set of the most likely hypotheses is first determined on the basis of $\Pr(DA \mid \text{history})$ and then rescored by $\Pr(DA \mid \text{words})$. These last two strategies only differ for the criterion with which the size of the restricted set is chosen. In one case, referred to as N-best in Table 6, the size of this set is defined independently from the input. On the contrary, in the second case this size is decided dynamically, and depends on the input: all and only the hypotheses whose probability difference from the best one is less than a threshold δ are chosen for rescoring.

The parameters used in the integration, i.e. the linear combination factor, N and δ, were chosen using a trial-and-error approach. Results obtained using single knowledge sources (SCT and DA trigrams) and the three above described types of integration of SCT and DA trigram statistics, are reported in Table 6. As usual, the experiments have been performed in LLO.

The experimental results show that the best strategy is the one called δ-best rescoring, in which the size of the set of hypotheses to be rescored is dynamically chosen. Nevertheless, the difference in performance between the first and the third strategies is slight, while all of them show an improvement with respect to the classifier simply based on segment words.

It is worth noting that all experiments are in some measure optimistic, as transcriptions are used instead of the recognizer outputs and $\Pr(DA \mid \text{history})$ was computed on the true history.

6. CONCLUSIONS AND FUTURE WORK

Even if the results obtained so far are encouraging, a lot of work is planned that is expected to improve performance. Where acoustic recognizer is concerned, the next step will regard the more detailed modelinig of extra-linguistic phenomena.

At present, most of the efforts concerning the understanding module have been devoted to SCTs design. Nevertheless, other parts need to be improved, including the preprocessing, which should deal with local phenomena. If a syntactical score is produced, this information could be integrated with that yet to be considered to focus the search for input semantic representation.

Depending on SCT performance, more or less effort has to be devoted to the development of the knowledge-based mechanism that completes the understanding module. Moreover, the dialogue history was taken into account in the more direct and simple way. A more sophisticated dialogue description is expected to help the classification.

On the other hand, all improvements are conditioned on the availability of a sufficient amount of data. Therefore, new data acquisitions are being performed, and further ones are scheduled for the future.

REFERENCES

[1] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo. Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1391–1394, Yokohama, Japan, 1994.

[2] B. Angelini, M. Cettolo, A. Corazza, D. Falavigna, A. Lavelli, G. Lazzari, F. Pianesi, and O. Stock. Preliminary Results of the C-STAR Project at IRST. In *Proc. of C-STAR II 96: ATR International Workshop on Speech Translation*, Kyoto, Japan, 1996.

[3] F. Brugnara and M. Cettolo. Improvements in Tree-based Language Model Representation. In *Proceedings of the 4th European Conference on Speech Communication and Technology*, pages 1797–1800, Madrid, Spain, 1995.

[4] M. Cettolo, A. Corazza, and R. De Mori. A Mixed Approach to Speech Understanding. In *Proc. of ICSLP*, Philadelphia, USA, 1996.

[5] R. De Mori and R. Kuhn. The Application of Semantic Classification Trees to Natural Language Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-17(5):449–460, May 1995.

[6] M. Federico, M. Cettolo, F. Brugnara, and G. Antoniol. Language Modeling for Efficient Beam-Search. *Computer Speech and Language*, 9(4):353–379, 1995.

[7] B. Suhm, L. Levin, N. Coccaro, and *al.* Speech-Language Integration in a Multi-Lingual Speech Translation System. In *ATR Collection 04*, pages 73–79. 1994.

[8] A. Waibel, M. Finke, D. Gates, and *al.* JANUS-II – Translation of Spontaneous Conversational Speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages I:409–412, Atlanta, Georgia, USA, 1996.