END-TO-END EVALUATION IN VERBMOBIL I

Rita Nübel
IAI
Martin-Luther-Straße 14
66111 Saarbrücken
Germany
rita@iai.uni-sb.de

Abstract

VERBMOBIL is a speech-to-speech translation system for spoken dialogues between two speakers. The application scenario is appointment scheduling for business meetings, with spoken dialogues between two speakers. Both dialogue participants have at least a passive knowledge of English which serves as intermediate language[1]. The transfer directions are German to English and Japanese to English. A special feature of VERBMOBIL is that translations are produced on demand when the dialogue participants are unable to express themselves in English and therefore prefer to use their mother tongue.

In this paper[2] we present the criteria and the evaluation procedure for evaluating the translation quality of the VERBMOBIL prototype. The evaluated data have been produced by three concurrent processing methods that are integrated in the VERBMOBIL prototype. These processing methods differ with respect to processing depth, processing speed and translation quality ([2], p. 2). The paper is structured as follows: we start by giving a short description of the VERBMOBIL architecture focusing on the concurrent linguistic analyses and transfer processes which lead to three alternative translation outputs for each turn[3]. In section two we outline the evaluation procedure and criteria. The third section discusses the evaluation results, and the conclusion of the paper gives an outlook to future applications of automated evaluation procedures for machine translation (MT) based on an MT architecture where several concurrent translation approaches are integrated.

1 Introduction

VERBMOBIL is a speech-to-speech translation system for spontaneous face-to-face dialogues. The VERBMOBIL domain is restricted to the negotiation of business meetings, but it is planned to extend the domain to travel planning. The VERBMOBIL system translates from German and Japanese into English. It is assumed that the dialogue partners have at least a passive knowledge of English which serves as intermediate language. The system is composed of a number of components such as syntactic-semantic analysis, transfer, generation, semantic evaluation, dialogue processing and front and back end components like speech recognition and synthesis. The communication between these components is controlled by a number of technical modules, which coordinate the interaction of more than 40 functional modules.

Two concurrent concepts of linguistic analysis have been implemented[4], namely the deep and shallow analysis. Deep processing is performed by a syntax-semantics component, a transfer and a target

[1] See [2] for a more detailed description of the objectives of VERBMOBIL.

[2] I would like to thank Ute Hauck for her assistance with Word and Visio. Special thanks to Elisabeth Maier for suggestions and comments which helped to improve this paper. The responsibility for the contents of the paper lies with the author.

[3] A turn comprises a speaker's contribution within the dialogue at a given point and may range from single-word utterances like "hello" up to several sentences.

[4] See fig. 1

generation component. These components interact with the modules semantic evaluation and dialog processing which provide additional information, e.g. for disambiguation purposes. The output of the unification-based semantic construction, which takes as input a word lattice (which in turn is the output of the acoustic analysis) is mapped onto a canonical representation called VERBMOBIL Interface Term (VIT [3]). The transfer output is also represented as a VIT, which is further processed by the English generation component. In a last step, the English synthesis produces the speech signals of the resulting translations.

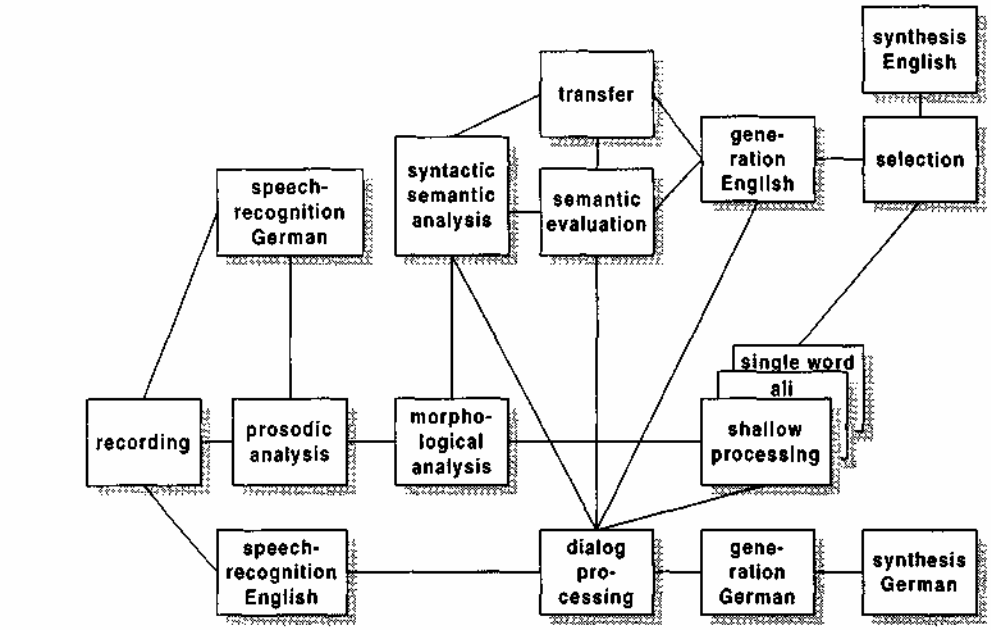

Figure 1: The VERBMOBIL Architecture[5]

The shallow analysis comprises two approaches to translation, namely translations based on dialogue acts[6], and translations based on the example-based approach. The first approach relies on the correct identification of dialogue acts underlying a turn which trigger template-based translations of dialogue act information. Propositional elements such as proper names and time expressions are integrated into the template and are then translated. The latter approach called ALI produces so-called schematic or example-based translations. The database on which these translations are based contains about 20.000 entries ([2], p. 2). Both shallow approaches work in parallel together with the deep analysis. This means that in the ideal case there are three possible translations available for each turn. The selection of the translation to be synthesized is made by an elementary version of the selection module. In the VERBMOBIL prototype, this module uses simple heuristics such as system performance, i.e. in case the deep analysis does not produce an output, the translation of the shallow processing is chosen (here, the dialogue act-based output would have priority over the ALI-output).

2 Evaluation Procedure and Criteria

For the final evaluation of the first VERBMOBIL prototype we examined about 20.000 turns which came from a set of recorded dialogue data ([2], p. 4). The transliterations of the original audio data together with their translations were evaluated by teams of professional translators and linguists via the Internet

[5] See [2], p. 3

[6] In VERBMOBIL, a dialogue act conveys the illocutionary and partially also the propositional information of an utterance, see [1] for a more detailed description

since the evaluators group was geographically distributed. In earlier evaluation rounds, the translations had also been evaluated by lay interpreters. The reason behind this strategy of looking at translations from different perspectives was to find out whether the potential user of VERBMOBIL would more easily be inclined to accept a bad translation than a professional interpreter, who would never use VERBMOBIL for obvious reasons. This was indeed the case. This parallel approach was given up in the final evaluation for the sake of statistic simplicity. What was required from the final evaluation were clear-cut "acceptable/non-acceptable" decisions based on judgments by a homogeneous group of evaluators. It has to be pointed out that the evaluators did not see the output of the speech recognition. The definition of the end-to-end evaluation implied that the translation quality should be assessed no matter what kind of defective input the linguistic components had to tackle.

2.1 Evaluators' Profile

The final VERBMOBIL evaluation was carried out by professional interpreters and translators as well as by linguists. All evaluators were native speakers of German. No special training for the evaluation task was given to them, but since the group was geographically distributed over different sites, a continuous information exchange took place in order to avoid rating divergence. The preservation of rating consistency was the most crucial aspect of the evaluation as a whole. Rating divergence naturally resulted from the evaluators' professional background and the well-known problem of lack of objective measurement of the output quality [7]. The linguists tended to accept stylistic oddities more often than the interpreters whereas the latter group were more apt to license grammatical errors (1, 2):

(1) *Wann machen wir die Besprechung dafür*
When are we doing a meeting for that?

Linguist: NOT ACCEPTABLE
Interpreter: ACCEPTABLE

(2) *Wir suchen uns Montag dafür aus*
How about meeting, say Monday?

Linguist: ACCEPTABLE
Interpreter: NOT ACCEPTABLE

However, in the overall picture the rating divergence did not lead to unreliable results: this was checked by comparing the ratings of both groups for identical test sets which diverged in less than 1 % of all evaluated turns.

2.2 Evaluation Procedure

While public test data is used in the development cycles, the end-to-end evaluation has been carried out with far more than 70% unseen test data. For the sake of objectivity, the data is presented to the evaluators without explicitly marking unseen and development data, such that no "expectations" as to the system's translation performance influence the rating. The translation results are presented together with the transliterations of the audio data. The evaluation units consist of a whole turn. The evaluation does not differentiate between very long or very short turns, i.e. there is no weighting of the length of a given turn, as was for instance done in the JANUS evaluation [5]. Since a detailed data analysis has shown that long turns do not necessarily constitute a significant source of errors, it is possible to do without the segmentation of turns into smaller semantic units. Additionally, we believe that splitting turns into smaller units which are then evaluated separately is not responsive to the evaluation as a true end-to-end evaluation task. In the scenario of spoken dialogues a transfer unit consists of the turn as *a whole* and should therefore be assessed as such. The alternative translation results obtained from the concurrent

analyses are presented anonymously, so that the evaluators do not know which processing method a translation came from. Once again, the evaluators should not be biased by presupposed requirements on the different approaches: it may be expected that high-quality translation is produced by deep processing whereas low-quality output of the shallow processing is acceptable as well.

A test file consists of 60 to 66 turns which in almost all cases constitute a coherent dialogue or at least a dialogue fragment. This is an important requirement for the presentation of the test data. There are a few cases where a test file is arbitrarily composed of a number of turns, which give rise to different translations due to the lacking dialogue context (3):

(3) *Geht das da nicht*
Module 1: *That doesn't suit me.*
Module 2: *Isn't it possible?*

In (3), the different processing methods produced completely different outputs. This is on the one hand due to the isolated presentation of this turn. Its context could not be deduced from preceding turns, and on the other hand, an unambiguous interpretation is also impossible by a lack of punctuation signs. In this case we decided to accept both translations although it is clear that in a real discourse situation only one of the translations is correct.

2.3 Evaluation Criteria

The evaluation task consists of an assessment of the "approximate correctness" of the given translation. Of course, the ultimate subjectivity of this quality measure imposes a certain degree of subjectivity on the evaluation criteria as well. The subjectivity of MT evaluations carried out by humans is a widely discussed problem [4], and no overall solution has been developed so far. We tried to relativize the subjectivity of grading by combining objectively measurable criteria with the more subjective ones. The most important requirement on the evaluation criteria is their responsiveness to the task-specificity of VERBMOBIL. We argue that VERBMOBIL, like most machine translation systems, has to be regarded as a tool to fulfill a specific task in a specific application domain[7]. The definition is in line with the concept of so-called "adequacy evaluations", which assess a system (as a whole, not just its output quality) with respect to its actual use in a specific setting, the evaluation thus being centered around the user requirements. The criteria we applied to measure the approximate correctness of a translation are the following:

1) **Preservation of the information content**
   Information preservation means that the relevant information, i.e. illocutionary and propositional content has to be conveyed in the translation. Stress was put on the correct translation of all kinds of time expressions, since the negotiation of time budgets is central to the VERBMOBIL application domain. In case one or more time expressions within a turn are not translated correctly, the translation is assigned the grade *unacceptable.*
2) **Understandability (comprehensibility) of the translation**
   The criterion of understandability is not objectively measurable and is considered as relative criterion.

The criteria are not weighed, i.e. no criterion has priority over the other, but both of them have to be fulfilled for a translation to be rated as *acceptable* translation. For this evaluation we applied a simple yes/no rating. This means that we have not developed a fine-grained rating scale, which would have

[7] See [6] for the discussion of the task-orientedness of MT.

blurred the statistics and which would have resulted in quite vague categories varying from e.g. a 'high degree' to a 'middle degree', a 'lower degree', etc. of approximate correctness. As opposed to the JANUS evaluation, the translated turns or parts of them have not been classified as belonging to the VERBMOBIL domain or not. Whereas in the JANUS evaluation, this classification triggers different treatments of utterance fragments depending on their domain status, we favor the evaluation of a turn as a whole and with respect to the overall translation task. Since no classification of utterances with respect to their in-domain or out-of-domain status has been defined for the VERBMOBIL translation task *before* the evaluation took place, we have restrained from further subdividing the translation task which would manipulate the results in favor of a better outcome. The only restriction on the input is that the utterances must contain words of the official VERBMOBIL wordlist (2461 words).

3 Evaluation Results

The results of the final end-to-end evaluation of the first VERBMOBIL prototype are shown in figure (2) below:

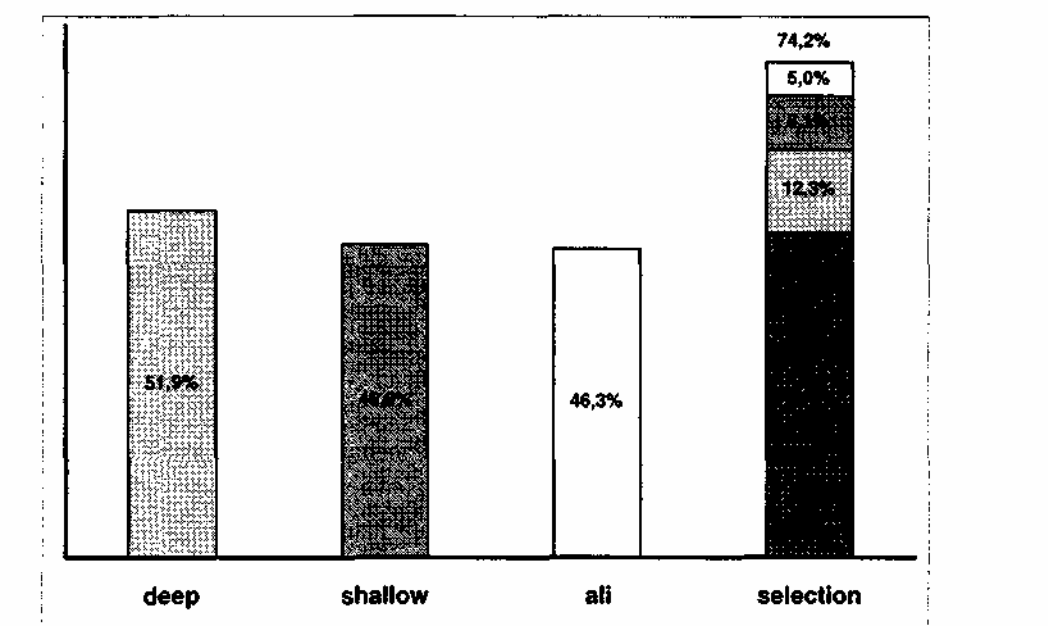

Figure 2: Evaluation Results

As expected, deep analysis achieves the best results. The results of the two shallow approaches are nearly identical. The overall translation rate is 74.2% which means that the concurrent approaches have complementary strengths. We examined the data to find out why one approach is constantly more successful than the others. Both shallow processing methods obviously are able to deal with at least a restricted amount of the data the deep processing cannot handle appropriately. This explains the complementary strength of all three approaches. Shallow processing based on dialogue acts is generally more successful than deep processing and the example-based approach if the system has to cope with relatively long turns containing more than two utterances with more than 20 words. To keep processing time at a reasonable level, a time-out mechanism forces the system to stop the analysis during deep processing if no result is produced within a fixed time limit. Since the deep analysis of long turns is quite time-consuming, the processing is timed-out in many of these cases, where the fast shallow approach produces quite acceptable results (4, 5):

(4) *Also ich würde sagen da es sich um ein Arbeitsfrühstück handelt, da würden wir uns vielleicht doch eher früher treffen, gut, wenn Ihnen das recht wäre, dann würden wir uns am sechzehnten November treffen, okay*
Shallow: *Well, how about meeting, say earlier? When would it fit you? How about meeting the sixteenth of November?* ACCEPTABLE
ALI: *Well I would say that since it is a business breakfast we should perhaps meet earlier* NOT ACCEPTABLE
Deep: no_result NOT ACCEPTABLE

(5) *Äh, eine Woche später war das ja, was wir zuerst hatten denke ich, äh am achtzehnten bis zum zweiundzwanzigsten ginge es noch*
Shallow: *How about meeting then? How about meeting say the eighteenth till the twenty - second?* ACCEPTABLE
ALI: *And I one week later have to inform you that I'm booked on the 18th till the 22nd and .* NOT ACCEPTABLE
Deep: no_result NOT ACCEPTABLE

On the other hand, in most cases the strength of the example-based approach consists of the correct analysis and translation of idiomatic expressions which are not yet covered by deep processing. They also constitute a source of errors for the dialogue act-based shallow processing (6):

(6) *Darüber habe ich mir auch schon Gedanken gemacht*
Shallow: *that's fine with me.* NOT ACCEPTABLE
ALI: *I've already been thinking about that.* ACCEPTABLE
Deep: *I've already done for me ideas about that too.* NOT ACCEPTABLE

Finally, the obvious strength of the deep analysis lies in the correct handling of "typical" data, i.e. turns of average length (<20 words, i.e. max. 2 utterances) containing time expressions of all kinds (7):

(7) *Wie wäre es denn Dienstagnachmittag um fünfzehn Uhr fünfundvierzig nach der Vorlesung*
Shallow: *how about meeting then? How about meeting say Tuesday in the afternoon?* NOT ACCEPTABLE
ALI: *How about Tuesday eight in the afternoon at 3.45 after class?* NOT ACCEPTABLE
Deep: *how about on Tuesday in the afternoon at a quarter to four after the lecture?* ACCEPTABLE

(8) *Das Wochenende davor würde mir passen*
Shallow: *how about meeting say the weekend?* NOT ACCEPTABLE
ALI: *weekend would suit me.* NOT ACCEPTABLE
Deep: *The weekend before would suit me.* ACCEPTABLE

(9) *Ich komme an einem Montag*
Shallow: *well, Monday!* NOT ACCEPTABLE
ALI: *I'll come Monday then.* NOT ACCEPTABLE
Deep: *I will come on a Monday.* ACCEPTABLE

Although in many of these cases the shallow approaches did also produce acceptable output, the output of deep processing was chosen automatically by the selection component because of the principally better translation quality. In 12.3% of the testcases only the deep analysis produced acceptable output, in

8.1% only the dialogue act-based approach, and in 5.0% of the test cases only the example-based approach produced acceptable output.

4 Related Work

End-to-end evaluations of a speech-to-speech translation system have also been carried out in the context of the JANUS project [5]. Similar to VERBMOBIL, JANUS is a multi-lingual speech-to-speech translation system for Spanish, English, German, Japanese and Korean. The JANUS system achieved a translation rate of 58% for the more accurate translation module, and 66% for the more robust translation module ([5], p. 10). There are several differences between the end-to-end evaluations of VERBMOBIL and JANUS which we will describe in short: The most striking difference concerning the size of the test data is that the VERBMOBIL evaluation is based on more than 20.000 turns whereas the test data used for the JANUS evaluation consists of 349 utterances only. Another characteristic of the JANUS evaluation is that utterances or parts of utterances are classified according to their domain relevance: out-of-domain data are evaluated separately. Although one might argue in favor of a separation of data for the testing of individual modules to support a more precise localization of weaknesses at the end of development cycles, this strategy was not adopted for the VERBMOBIL evaluation of the overall translation performance.

Another difference between the two evaluations related to the concept of evaluation unit. In VERBMOBIL, an evaluation unit naturally corresponds to a complete turn. In JANUS, utterances are broken into semantically based chunks ([5], p. 5). These smaller units are then evaluated. The intention behind this method is to be more responsive to the domain-relevance classification and to the length of utterances. We restrained from this subclassification of data since the huge amount of VERBMOBIL evaluation data automatically guarantees a balance between very long and very short utterances as well as between different levels of domain relevance.

5 Conclusion

The evaluation criteria and the procedure used for the end-to-end evaluation of the VERBMOBIL prototype proved to be consistently applicable to very huge amounts of evaluation data which have been produced by concurrent transfer approaches.

The analysis of the evaluated data shows that the complementary strength of the different transfer approaches relates to characteristics of both the input data and to the specifications of the transfer approaches which can serve as more fine-grained criteria for the selection of the best of the parallel outputs. A third layer for the derivation of evaluation criteria refers to user expectations. This information has to be taken into account for future applications of automated selection mechanisms. They can be envisaged for MT architectures in a network environment like the Internet or within corporate intranets[8]. A prerequisite is the availability of various concurrent MT engines each of which is designed for a specialized translation task. In the ideal case, the MT engines are integrated (situated) in a complex authoring environment and fulfill a special subtask (i.e. translation) of a more complex task of multi-lingual text management. A selection component (or software agent) communicates with other software agents which support the communication with and between the MT engines. The selection component needs information concerning the translation task for the evaluation and selection of the appropriate MT engine. The definition of a user profile serves as a collection of criteria on which such a selection is based. A possible criterion is the determination of the required output quality, e.g. does the

[8] I would like to thank J. Schütz who introduced me to the notion of translation broking in a network environment and with whom I had many discussions which led to the central considerations concerning automated evaluation procedures for MT.

user want a raw translation or a high-quality translation or is the translation task concerned with multilingual retrieval and abstracting. Other criteria refer to the input text itself: has it been produced with or checked by (preferably integrated) authoring tools such as grammar and style checkers which are designed to match the specifications of one or more of the available MT components to increase the translation quality. Selection based on user profiling takes place before the text is translated and consists in pre-selecting the most appropriate MT component. Alternatively, the best translation results are selected after all concurrent MT components have performed the task. This procedure can be compared to a true evaluation. In this case, evaluation criteria relate to the individual contributions of the MT components (together with the information provided by the user profile). If for instance the user requires a high quality translation and if the selection component gets the information that the example-based MT component has achieved a 100% match, then the output produced by this component will receive the highest quality score. It should be noted that the results of the evaluation have to be considered as *relative* results due to the comparative method with which they have been achieved.

These ideas can be considered as first steps towards a more objective (and less time-consuming) automated assessment of the quality of machine translation output. Further investigations in the direction of more flexible MT architectures are required. They allow for the integration of various MT components with very specific tasks. The development and integration of tools to support the fulfillment of these tasks are also required for future MT applications and are also a prerequisite for the development of task-oriented automatic evaluation procedures.

References

[1] **Alexandersson, J., Reithinger, N., Maier, E.:** Insights into the Dialogue Processing in VERBMOBIL. In: *Proceedings of ANLP-97,* Washington 1997, forthcoming (1997).

[2] **Bub, Th., Wahlster, W., Waibel, A.: VERBMOBIL:** The Combination of Deep and Shallow Processing For Spontaneous Speech Translation. Paper Submitted to *ICASSP* (1997).

[3] **Dorna, M.:** The ADT-Package for the VERBMOBIL Interface Term. VERBMOBIL-Report Nr.104, Stuttgart (1996).

[4] **Falkedal, K.:** Evaluation Methods for Machine Translation Systems. An Historical Overview and Critical Account. *Draft Report,* ISSCO, University of Geneva (1993).

[5] **Gates, D., Lavie, A. Levin, L., Waibel, A., Gavalda, M., Mayfield, L. Woszczyna, M.:** End-to-end Evaluation in JANUS: A Speech-to-speech Translation System. In: *Proceedings of the 6th European Conference on Artificial Intelligence (ECAI),* Budapest (1996).

[6] **Kay, M.:** *The Proper Place of Man and Machines in Language Translation.* Xerox Palo Alto Research Center, Palo Alto, Ca (1980).

[7] **King, M.:** Evaluating Natural Language Processing Systems. In: *Communications of the ACM,* January 1996/Vol.39, No.l (1996).

[8] **Schütz, J.:** Network-based Machine Translation Services. In: *Proceedings of the EAMT Workshop, TKE'96,* Vienna, 29 - 30 August (1996).

[9] **Sparck Jones, K., Galliers, J.R.:** Evaluating Natural Language Processing Systems. Springer (1995).

[10] **White, J.S., O'Connell, T., O'Mara, F.E.:** The ARPA MT Evaluation Methodologies: Evolution, Lessons and Further Approaches. In: *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA),* Columbia (Md.) (1994).