From Human Evaluation to Automatic Selection of Good Translations

Damir Ćavar, Uwe Küssner, Dan Tidhar

Technische Universität Berlin, FB Informatik
Sekr. Fr 6-10, Franklinstr. 28/29, 10587 Berlin, Germany
{cavar, uk, dan}@cs.tu-berlin.de

Abstract

In this paper we describe a machine learning method tailored to overcome the difficulty of selecting and putting together translated segments in the *Verbmobil* system. We use off line human feedback to determine an optimized confidence rescaling scheme for the confidence values provided by four independent and competing translation paths in *Verbmobil.*

1. Introduction

Within the machine translation system *Verbmobil* (Wahlster, 1993), translations of spoken input are performed simultaneously in four independent translation paths. Each translation path implements a completely different MT strategy. The different translation paths incrementally generate translation segments that are combined to one output translation by a selection procedure. The task of the selection procedure is to find the best translation of the single segments and generate a translation output, which is the optimal set of translated segments.

The selection procedure relies on confidence values that are delivered together with the translated segments from each of the alternative translation components. Since the confidence values are computed by independent components that are based on fundamentally different MT strategies, they are not directly comparable, neither with each other, nor with the objective judgments of human evaluators, i.e. they need to be rescaled in order to gain comparative significance. In this paper, we describe the strategy we have explored in order to acquire the necessary human annotations of the translation quality, that are used as the bootstrapping data for an optimized confidence rescaling schema.

For our purpose, the results from human evaluation are the key data. In order to guarantee maximal reliability we make use of different strategies developed and used for example in experimental psychology. The annotation task itself has to be designed in a way to resolve optimally the tension between the need to be maximally easy for the evaluators (low time resources, low cognitive effort) and maximally reliable and usable for the developers. The tasks for the evaluator were set up to consist of the following primitives: a. simple reading task, b. binary decision task, c. simple counting task, and the possibility to make notes. As is well known from experimental psychology and psycholinguistics, simple binary decision tasks (e.g. yes/no questions), for example, are answered much faster, and more reliably across items and across subjects by the evaluators than decision tasks that provide a decision scale. Counting tasks are used, both, to generate relevant annotated data (e.g. the number of relevant information units), and to provide automatic means for checking the evaluators reliability, given that some counting tasks can be performed automatically. The evaluator, for example, is randomly asked to count the number of words in either the input or the output, and the instructor explains that this is relevant for the evaluation. The number of mistakes the evaluator makes can be used to relativize the other evaluation results automatically. Such evaluation strategies and precise instructions, as we experienced, give very robust results.

Based on this evaluation, a set of 'off line' confidence values is calculated, and a list of alternative segment combinations is produced, sorted according to their corresponding off line confidence values. The annotators then process these lists in a second annotation phase, in which they are requested to select from each list a minimal subset of 'best' translations. The results of this second annotation round are then combined with the original 'on line' confidence values to form inequalities that express the annotators' preferences as a set of constraints on the linear rescaling coefficients.

2. The Various Translation Paths

The *Verbmobil* system includes four independent translations paths that operate in parallel. The input shared by all paths consists of sequences of annotated *Word Hypotheses Graphs* (WHG). Each *WHG* is produced by a speaker independent voice recognition module, and is annotated with additional prosodic information and pause information by a prosody module (Buckow et al, 1998). In principle, every translation subsystem chooses independently a path through the *WHG,* and a possible segmentation according to its grammar and to the prosody module information. This implies that even though all

translation paths are sharing the same input data structure, both the chosen input string and its chosen segmentation may well be different for each path. In this section we provide the reader with very brief descriptions of the different translation subsystems, along with their respective methods for calculating confidence values.

The *ali* subsystem implements an example based translation approach. Confidence values are calculated according to the matching-level of the input string with its counterparts in the database.

The *stattrans* (Och et al, 1999) subsystem is a statistical translation system. Confidence values are calculated according to a statistical language model of the target language, in conjunction with a statistical translation model.

The *syndialog* (Kipp et al, 1999) subsystem is a dialogue act based translation system. Here the translation invariant consists of a recognized dialogue act, together with its extracted propositional content. The confidence value reflects the probability that the dialogue act was recognized correctly, together with the extent to which the propositional content was successfully extracted.

The *deep* translation path in itself consists of multiple pipelined modules: linguistic analysis, semantic construction, dialogue and discourse semantics, and transfer (Emele et al, 1996) and generation (Kilger et al, 1995) components. The transfer module is supported with disambiguation information by the context (Koch et al, 2000) and dialogue modules. The linguistic analysis part consists of several parsers which, in rum, also operate in parallel (Ruland et al, 1998). They include an HPSG parser, a Chunk Parser and a statistical parser, all producing data structures of the same kind, namely, the *Verbmobil Interface Terms* (VITs) (Dorna, 1999). Thus, within the deep processing path, a selection problem arises, similar to the larger scale problem of selecting the best translation. This internal selection process within the deep path is based on a probabilistic *VIT* model. Confidence values within the deep path are computed according to the amount of coverage of the input string by the selected parse, and are subject to modifications as a byproduct of combining and repairing rales that operate within the semantics mechanism. Another source of information which is used for calculating the 'deep' confidence values is the generation module, which estimates the percentage of each transferred *VIT* which can be successfully realized in the target language.

Although all confidence values are finally scaled to the interval [0,100] by their respective generating modules, there seems to be hardly any reason to believe that such fundamentally different calculation methods would yield magnitudes that are directly comparable with one another. As expected, our experience has shown that when confidence values are taken as such, without any further modification, their comparative significance is indeed very limited.

3. The Selection Procedure

For the language pairs English-German and German-English, *Verbmobil* applies four different translation methods that operate in parallel, according to four alternative approaches to machine translation, thus increasing the system's robustness and versatility. Since the system should always produce exactly one translation for each input utterance that it encounters, a selection procedure is necessary, which would choose the best alternative for each given utterance.

In order to benefit more from this diversity of translation methods, the alternative translations are furthermore combined within the boundaries of single utterances, so as to form a new compound translation.

Each translation module calculates a confidence value for each of the translations that it produces, to serve as a guiding criterion for the selection procedure. However, since the various translation methods are fundamentally different from one another, the resulting confidence values cannot be compared per se. Whereas we do assume a general correspondence between confidence values and translation quality within each one of the modules, there is no guaranty whatsoever that a high value delivered by a certain module would indeed signify a better translation when compared with another value, even a much lower one, which was delivered by another module. An additional step needs to be taken in order to make the confidence values comparable with one another.

An important presupposition that has been adopted throughout the current work is that the desirable rescaling can be well approximated by means of linear polynomials. The computational benefits of this assumption are immense, as it allows us to remain within the relatively friendly realm of linear equations (albeit inconsistent). The price that we have to pay in terms of precision is not as big as one might expect, because the crucial matter to our case is the comparative behavior of the obtained confidence curves, i.e. the breakpoints in which one overtakes the other, rather than the precise details of their behavior in between.

Once the rescaling coefficients have been determined by the learning procedure, the selection procedure can be performed as follows: for each given utterance, all possible translated segment sequences that combine to a full translation are generated. Their respective normalized confidence values are then calculated by applying the linear rescaling coefficients, and then integrating with respect to the time axis, in order to favor sequences with better source utterance coverage. The best sequence can then be chosen according to the normalized confidence values. It should be noted that not all sequences need to be actually generated and tested, due to the incorporation of Dijkstra's well known "Shortest Path" algorithm (e.g. in Cormen et al 989).

4. The Learning Procedure

Learning the rescaling coefficients is performed off line, and should normally take place only once, unless new training data is assembled, or new criteria for the desirable system behaviour have been formulated. The learning cycle consists of incorporating human feedback (training set annotation) and finding a set of rescaling coefficients so as to yield a selection procedure with optimal or close to optimal accord with the human annotations. The first step in the learning procedure is choosing the set of training data. This choice has a direct influence on the learning's result, and, of course, on the amount of time and resources that it requires. In the course of our work we've performed this procedure several times, with training sets of various sizes, all taken from a corpus of test dialogues, designed to provide a reasonable coverage of the desirable functionality of the current {\em Verbmobil} version.

Since the optimization algorithm (described below) normally terminates within no more than a couple of hours, the main bottle neck in terms of time consumption have normally been the human annotators. With what appears to be, from our experience, a reasonably large training set, i.e. a set of 7 from the above mentioned test dialogues (including 240 dialogue turns and 1980 different segments), the complete learning cycle can be performed within a few days, depending on the annotators' diligence, of course. Once a training set has been determined, it is first fed through the system, while separately storing the outputs produced by the various translation modules.

The system's output is then subject to two phases of annotation, resulting in a uniquely determined 'best' sequence of translated segments for each input utterance. The next task is to learn the appropriate linear rescaling, that would maximize the accord between the new, rescaled confidence values, and the preferences dictated by the newly given 'best' sequences. In order to do that, we first generate a large set of inequalities, and then obtain their optimal, or close to optimal solution.

The two annotation phases can be described as follows: first, the outputs of the alternative translations paths are annotated separately, so as to enable the calculation of the 'off line confidence values' as described below. For each dialogue turn, all possible combinations of translated segments that cover the input are then generated. For each of those possible combinations, an overall off line confidence value is calculated, in a similar way to which the 'online' confidence is calculated, leaving out the rescaling coefficients, but keeping the time axis integration.

These segment combinations are then presented to the annotators for a second round, sorted according to their respective off line confidence values. The annotator is requested at this stage merely to select the best segment combination, which would normally be one of the first to appear on the list.

The first annotation stage may be described as 'theory assisted annotation', and the second is its more intuitive complement. To assist the first annotation round we have compiled a set of annotation criteria, and designed a specialized annotation tool for their application.

These criteria direct the annotator's attention to 'essential information items', and refer to the number of such items that have been deleted, inserted or maintained during the translation. Other criteria are the semantic and syntactic correctness of the translated utterance as well as those of the source utterance. The separate annotation of these criteria allows us to express the 'off line confidence' as their weighted linear combination. The different weights can be seen as implicitly establishing a method of quantifying translation quality. One can determine, for instance, which is of higher importance - syntactical correctness, or the transmission of all essential information items. Using the vague notion of 'translation quality' as a single criterion would have definitely caused a great divergence in personal annotation style and preferences, as can be very well exemplified by the case of the dialogue act based translation: some people find word by word correctness of a translation much more important than the dialogue act invariance, while others argue exactly the opposite (Schmitz, 1997), (Schmitz et al, 1995).

Once the best segment sequences for each utterance have been determined by the completed annotation procedure, a set of inequalities is created using the linear rescaling coefficients as variables. This is done simply by stating the requirement that the normalized confidence value of the best segment sequence should be better than the normalized confidence values of each one of the other possible sequences. For each utterance with n possible segment sequences, this requirement is expressed by $(n-1)$ inequalities.

It is worth mentioning at this point that it sometimes occurs during the second annotation phase, that numerous sequences relating to the same utterance are considered 'equally best' by the annotator. In such cases, when not all sequences are concerned but only a subset of all possible sequences, we have allowed the annotator to select multiple sequences as 'best', correspondingly multiplying the number of inequalities that are introduced by the utterance in question. These multiple sets are known in advance to be inconsistent, as they in fact formulate contradictory requirements. Since the optimisation procedure attempts to satisfy the largest possible subset of inequalities, the logical relation between such contradicting sets can be seen as disjunction rather than conjunction, and they do seem to contribute to the learning process, because the different 'equally best' sequences are still favoured in comparison to all other sequences relating to the same utterance.

The overall resulting set of inequalities is normally very large, and can be expected to be consistent only in

a very idealized world, even in the absence of 'equally best' annotations. The inconsistencies reflect many imperfections that characterize both the problem at hand and the long way to its solution, most outstanding of which is the fact that the original confidence values, as useful as they may be, are nevertheless far from reflecting the human annotation and evaluation results, which are, furthermore, not always consistent among themselves.

The rest of the learning process consists in trying to satisfy as many inequalities as possible without reaching a contradiction.

The problem of finding the best rescaling coefficients reduces itself, under the above mentioned presuppositions, to that of finding the maximal consistent subset of inequalities within a larger, most likely inconsistent, set of linear inequalities, and solving it. In (Amaldi et al, 1997), the problem of extracting close-to-maximum consistent subsystems from an inconsistent linear system (MAX CS) is treated as part of a strategy for solving the problem of partitioning an inconsistent linear system into a minimal number of consistent subsystems (MIN PCS).

Both problems are NP-hard, but through a thermal variation of previous work by (Agmon, 1954) and (Motzkin et al, 1954), a greedy algorithm is formulated by (Amaldi et al 1997), which can serve as an effective heuristic for obtaining optimal or near to optimal solutions for MAX CS. Implementing this algorithm in the C language enabled us to complete the learning cycle by finding a set of coefficients that maximizes, or at least nearly maximizes, the accord of the rescaled confidence values with the judgment provided by human annotators.

5. Conclusion

We have described certain difficulties that arise during the attempt to integrate multiple alternative translation paths and to choose their optimal combination into one 'best' translation. Using confidence values that originate from different translation modules as our basic selection criteria, we have introduced a learning method which enables us to perform the selection in close to maximal accord with decisions taken by human annotators. Along the way, we have also tackled the problematic aspects of translation evaluation as such, and described some additional sources of information that are used within our selection module. The extent to which this module succeeds in creating higher quality compound translations is of course highly dependent on the appropriate assignment of confidence values, which is performed by the various translation modules themselves.

Despite the relative simplicity of the methods that are currently being used by these modules for confidence calculation as such, applying our approach within the *Verbmobil* system has already yielded a significant improvement. The most recent *Verbmobil* evaluation results demonstrate this improvement very clearly. The evaluation is based on annotating five alternative translations for a chosen set of dialogue-turns. The translations provided by the four single translation paths, and the combined translation delivered by the selection module, were all marked by the annotators as 'good', 'intermediate', or 'bad'. Judged by the percentage of 'good' turns from the overall number of annotated turns, the selection module shows an improvement of 27.8% compared to the best result achieved by any single module.

6. References

Agmon S., 1954. The relaxation method for linear inequalities *Canadian Journal of Mathematics, 6:382-392.*

Amaldi E., Mattavelli M., 1997. A combinatorial optimization approach to extract piecewise linear structure from nonlinear data and an application to optical flow segmentation *TR 97-12, Cornell Computational Optimization Project, Cornell University, Ithaca NY, USA.*

Buckow J., Batliner A., Gallwitz F., Huber R., Nöth E., Warnke V., and Niemann H., 1998. Dovetailing of Acoustics and Prosody in Spontaneous Speech Recognition *Proc. Int. Conf. on Spoken Language Processing, volume 3, pages 571-574, Sydney, Australia.*

Cormen T., Leiserson C., Rivet L., 1989. Introduction to Algorithms *MIT Press, Cambridge, Massachusetts.*

Dorna M, 1999. The ADT Package for the Verbmobil Interface Term *Universität Stuttgart. Verbmobil Report 104X.*

Emele M., Dorna M., 1996. Efficient Implementation of a Semantic-based Transfer Approach *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI-96).*

Kilger A., Finkler W., 1995. Incremental Generation for Real-Time Applications *DFKI Report RR-95-11, German Research Center for Artificial Intelligence - DFKI GmbH.*

Kipp M., Alexandersson J., Reithinger N., 1999. Understanding Spontaneous Negotiation Dialogue *Proceedings of the IJCAI Workshop Knowledge and Reasoning in Practical Dialogue Systems, Stockholm, Sweden*

Koch S., Küssner U, Stede M, Tidhar D., 2000. Contextual reasoning in speech-to-speech translation

Proceedings of 2nd International Conference on Natural Language Processing NLP2000, Springer LNAI.

Motzkin T.S., Schoenberg I.J., 1954. The relaxation method for linear inequalities *Canadian Journal of Mathematics, 6:393-404.*

Och F.J., Tillmann C., Ney H., 1999. Improved Alignment models for Statistical Machine Translation *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora, University of Maryland.*

Ruland T., Rupp C.J., Spilker J., Weber H., Worm C, 1998. Making the Most of Multiplicity: A Multi-Parser Multi-Strategy Architecture for the Robust Processing of Spoken Language *Proceedings of ICSLP.*

Schmitz B., 1997. Pragmatikbasiertes Maschinelles Dolmetschen. *Dissertation, FB Informatik, TU Berlin.*

Schmitz B., Quantz J.J., 1995. Dialogue Acts in Automatic Dialogue Interpreting *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95), Leuven.*

Wahlster W., 1993. Verbmobil: Translation of face-to-face dialogues. *Proceedings of the Third European Conference of Speech Communication and Technology, Berlin.*