Learning to Select a Good Translation

Dan Tidhar and **Uwe Küssner**
Technische Universität Berlin
Fachbereich Informatik
Franklinstr. 28/29
D-10587 Berlin
Germany
{uk|dan}@cs.tu-berlin.de

Abstract

Within the machine translation system *Verbmobil*, translation is performed simultaneously by four independent translation modules. The four competing translations are combined by a selection module so as to form a single optimal output for each input utterance. The selection module relies on confidence values that are delivered together with each of the alternative translations. Since the confidence values are computed by four independent modules that are fundamentally different from one another, they are not directly comparable and need to be rescaled in order to gain comparative significance. In this paper we describe a machine learning method tailored to overcome this difficulty by using off-line human feedback to determine an appropriate confidence rescaling scheme. Additionally, we describe some other sources of information that are used for selecting between the competing translations, and describe the way in which the selection process relates to quality of service specifications.

1 Introduction

Verbmobil (Wahlster, 2000) is a speech to speech machine translation system, aimed at handling a wide range of spontaneous speech phenomena within the restricted domain of travel planning and appointment scheduling dialogues. For the language pairs English-German and German-English, four different translation methods are applied in parallel, thus increasing the system's robustness and versatility. Since exactly one translation should be produced for each input utterance, a selection procedure is necessary. In order to benefit more from this diversity, the alternative translations are furthermore combined within the boundaries of single utterances, so as to form new compound translations. Combining translations from different sources within a multi-thread MT system has already proved beneficial in the past (Frederking and Nirenburg, 1994). Our present work differs from the work reported in there in several ways (apart from the trivial fact that we use 'four heads' rather than three). Firstly, we attempt to investigate a systematic solution to the incomparability of the various confidence values. Secondly, as we deal with speech to speech rather than text to text translation, different segmentations for each given input string are allowed, making the segment combination process significantly more complicated.

1.1 Incomparability

Each translation module calculates a confidence value for each translation that it produces. However, since the various translation methods are fundamentally different from one another, the resulting confidence values cannot be directly compared across modules. 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.

1.2 Working Hypotheses

It should be noted that one of our working hypotheses, namely, that confidence values do generally reflect translation quality, also compensates to a certain extent for the lack of a wide range theory of translation, according to which translations of different sorts could be unanimously evaluated. The task of evaluating

translation quality is non-trivial also for human annotators, since the applicable criteria are diverse, and at the absence of a comprehensive translation theory, very often lead to contradicting conclusions. This difficulty is partially dealt with in section 4.1 below, but for practical reasons we tend to accept the need to rely on human judgment, partially theory assisted and partially intuitive, as inevitable. Another presupposition that underlies the current work is that the desirable rescaling can be well approximated by means of linear polynomials. This assumption allows us to remain within the relatively friendly realm of linear equations (albeit inconsistent), and reflects two basic guiding principles: firstly, that the rescaling is motivated by pragmatical needs, rather than by descriptive aspirations, and secondly, that it should not contradict the presupposed correlation between confidence and quality *within* each module, which implies that the rescaling functions should be monotonous.

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)*, produced by the speech recognizer. Each translation module chooses independently a path through the *WHG*, and a possible segmentation according to its grammar and to the prosody information (Buckow *et al.*, 1998). This implies that even though all translation modules share the same input data structure, both the chosen input string and its chosen segmentation may well differ across modules. This section provides the reader with very brief descriptions of the different translation subsystems, along with their respective confidence value calculation methods.

- 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 and Dorna, 1996) and generation (Kilger and Finkler, 1995) components. The transfer module receives disambiguation information from the context (Koch *et al.*, 2000) and dialogue modules. The linguistic analysis part consists of several parsers which, in turn, 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)* (Schiehlen *et al.*, 2000). 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 rules 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 transfered *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 dif-

ferent 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

In order to improve their comparative significance, the delivered confidence values $c(s)$, for each given segment s, are rescaled by linear functions of the form:

$$a \cdot c(s) + b \ . \tag{1}$$

Note that each input utterance is decomposed into several segments independently, and hence potentially differently, by each of the translation paths. The different segments are then combined to form a data structure which, by analogy to *Word Hypotheses Graph*, can be called *Translation Alternatives Graph (TAG)*. The size of this graph is bound by 4^n, which is reached if all translation paths happen to choose an identical partition into exactly n segments. The following vectorial notation was adopted in order to simplify the simultaneous reference to all translation paths. The linear coefficients are represented by the following four-dimensional vectors:

$$\vec{a} = \begin{pmatrix} a_{ali} \\ a_{syndialog} \\ a_{stattrans} \\ a_{deep} \end{pmatrix} \qquad \vec{b} = \begin{pmatrix} b_{ali} \\ b_{syndialog} \\ b_{stattrans} \\ b_{deep} \end{pmatrix} \ . \tag{2}$$

Single vector components can then be referred to by simple projections, if we represent the different translation paths as orthogonal unit vectors, so that $\vec{s}$ denotes the vector corresponding to the module by which s had been generated. The normalized confidence is then represented by:

$$(\vec{a} \cdot k(s) + \vec{b}) \cdot \vec{s} \ . \tag{3}$$

In order to express the desirable favoring of translations with higher input string coverage, the compared magnitudes are actually the (rescaled) confidence values integrated with respect to the time axis, rather than the (rescaled) confidence values as such. Let $\|s\|$ be the length of a segment s of the input stream, in milliseconds. Let **SEQ** be the set of all possible segment sequences within the TAG, and $Seq \in \mathbf{SEQ}$ any particular sequence.

We define the normalized confidence of Seq as follows:

$$C(Seq) = \sum_{s \in Seq} ((\vec{a}\ c(s) + \vec{b})\vec{s})\ \|s\|$$

This induces the following order relation:
$seq_1 \leq_C seq_2 \stackrel{def}{=} C(seq_1) \leq C(seq_2)$
Based on this relation, we define the set B of best sequences as follows:

$$B(\mathbf{SEQ}) = \{seq \in \mathbf{SEQ} \mid seq \text{ is a maximum element in } (\mathbf{SEQ}; \leq_C)\} \ . \tag{4}$$

The selection procedure consists in generating the various possible sequences, computing their respective normalized confidence values, and arbitrarily choosing a member of the set of best sequences. 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.*, 1989)).

4 The Learning Cycle

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 behavior 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 evaluation. A training set, consisting of test dialogues that cover the desirable system functionality, is fed through the system, while separately storing the outputs produced by the various translation modules. These are then subject to two phases of annotation (see section 4.1), resulting in a set of 'best' sequences of translated segments for each input utterance. The next task is to determine the appropriate linear rescaling, that would maximize the accord between the rescaled confidence values and the preferences expressed by those 'best' sequences. In order to do that, we first generate a large set of inequalities as described in section 4.2 below, and then approximate their optimal solution, as described in section 4.3.

4.1 Training Set Annotation

As mentioned above, evaluating alternative translations is a complex task, which sometimes appears to be difficult even for specially trained people. When one alternative seems highly appropriate and all the others are clearly wrong, a vigilant annotator would normally encounter very little difficulty. But when all options fall within the reasonable realm and differ only slightly from one another, or even more so, when all options are far from perfect, each having its uniquely combined weaknesses and advantages — what criterion should be used by the annotator to decide which weaknesses are more crucial than the others? Our human feedback cycle is twofold: 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 'on-line' confidence is calculated (see section 3), 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 and Quantz, 1995).

4.2 Generating Inequalities

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 optimization 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 favored 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.

4.3 Optimization Heuristics

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 and Mattavelli, 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 and Schoenberg, 1954), a greedy algorithm is formulated by (Amaldi and Mattavelli, 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 Additional Information Sources

Independently of the confidence rescaling process, we have made several attempts to incorporate additional information in order to refine the selection procedure. Some of these attempts, such as using probabilistic language model information, or inferring from the logical relation between the approximated propositional contents of neighboring utterances (e.g. trying to eliminate contradiction), have not been fruitful enough to be worth full description in the present work. The following two sections describe two attempts that do seem to be worth mentioning in further detail.

5.1 Dialogue Act Information

Our experience shows that the translation quality that is accomplished by the different modules varies, among the rest, according to the dialogue act at hand. This seems to be particularly true for **syndialog**, the dialogue act based translation path. Those dialogue acts that normally transmit very little propositional content, or those that transmit no propositional content at all, are normally handled better by **syndialog** compared to dialogue acts that transmit more information (such as INFORM, which can in principle transmit any proposition). The dialogue act recognition algorithm used by **syndialog** does not compute the single most likely dialog act, but rather a probability distribution of all possible dialogue acts[1] We represent the dialogue act probability distribution for a given segment s by the vector $\vec{da}(s)$, where each component denotes the conditional probability of a certain dialogue act, given the segment s:

$$\vec{da}(s) = \begin{pmatrix} P(suggest|s) \\ P(reject|s) \\ P(greet|s) \\ \vdots \end{pmatrix} . \qquad (5)$$

The vectors $\vec{a}$ and $\vec{b}$ from section 3 above are replaced by the matrices $\vec{A}$ and $\vec{B}$ which are simply a concatenation of the respective dialogue act vectors.
Let $\vec{A}^s = \vec{A} \cdot \vec{da}(s)$, and $\vec{B}^s = \vec{B} \cdot \vec{da}(s)$.
The normalized confidence value, with incorporated dialogue act information can then be expressed as:

$$C(Seq) = \sum_{s \in Seq} ((\vec{A}^s \cdot c(s) + \vec{B}^s) \cdot \vec{s}) \cdot \|s\| \ . \qquad (6)$$

5.2 Disambiguation Information

Within the **deep** translation path, several types of underspecification are used for representing ambiguities (Küssner, 1997), (Küssner, 1998), (Emele and Dorna, 1998). Whenever an ambiguity has to be resolved in order for the translation to succeed, resolution is triggered on demand (Buschbeck-Wolf, 1997). Several types

[1] For more information about dialogue acts in *Verbmobil*, see (Alexandersson *et al.*, 1997)

of disambiguation are performed by the context module (Koch *et al.*, 2000), which uses various knowledge sources in conjunction for resolving anaphorical and lexical ambiguities. Examples for such knowledge sources are world knowledge, knowledge about the dialogue state, as well as various sorts of morphological, syntactic and semantic information. Since the **deep** translation path is the only one that includes contextual disambiguation, its confidence value is incremented by the selection module whenever such ambiguities occur.

6 Quality of Service Parameters

Translation quality is perhaps the most significant Quality of Service (QoS) parameter as far as MT systems are concerned. The selection module and the learning procedure as described above, are indeed aimed at optimizing this parameter. Additionally, we have further experimented with our selection module in order to accommodate for other QoS parameters as well. Analogously to QoS in Open Distributed Programming (ODP), we can distinguish between the following main categories: timeliness, volume, and reliability. In the timeliness category, we refer to the delay from the beginning of the acoustic input till the beginning of the acoustic output, which is highly dependent on the system's incrementality. The algorithm described so far requires the presence of all translated segments within a given dialogue turn, before the selection itself can take place. This implies a relatively long delay, because the biggest possible increment unit, i.e. the whole turn, is being used. The maximal incrementality, and therefore the minimal delay, are achieved when the first ready segment is being chosen at each point. This implies, however, a possible deterioration in translation quality, and increasing the risk that due to segmentation differences across modules, no appropriate continuation would be found for the first segment that had been chosen. The latter is referred to as 'loss rate', and belongs to the reliability category of QoS dimensions. The trade-off between loss rate and incrementality is parameterized by the selection module, by selecting a segment as soon as n translation modules have delivered segments with similar segmentations ($1 \leq n \leq 4$). Within the volume category, we define the real time factor (RTF) as the relation between the overall processing time (from the beginning of acoustic input till the end of acoustic output) and the overall speaking time (beginning of acoustic input till the end of acoustic input). In order to support conformance to RTF specification for the translation service, the selection module supports a QoS signal interface. A QoS management module monitors the runtime behavior of the translation modules, and signals the selection process if the estimated RTF is expected to exceed the specification. Upon receiving such a signal, the selection module attempts to complete its output without waiting for further translated segments.

7 Conclusion

We have described certain difficulties that arise from 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 criterion, we have introduced a learning method which enables us to select in maximal accord with decisions taken by human annotators. Along the way, we have also tackled some problematic aspects of translation evaluation as such, described some additional sources of information that are used by our selection module, and briefly sketched the way in which it supports quality of service specifications. 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, performed by the translation modules themselves. As a rough criterion for evaluating our success, we compared the selection module's output to the best results achieved by a single translation path. Recent *Verbmobil* evaluation results demonstrate an improvement of 27.8% achieved by the selection module, measured by the number of dialogue turns that were marked 'good' by annotators who were presented with five alternative translations for each turn, namely, those delivered by the four single paths, and the compound translation delivered by the selection module.

References

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

J.Alexandersson, B.Buschbeck-Wolf, T.Fujinami, M.Kipp, S.Koch, E.Maier, N.Reithinger, B.Schmitz, M.Siegel. *Dialogue Acts in VERBMOBIL-2 Second Edition*, DFKI Saarbrücken, Universität Stuttgart, Technische Universität Berlin, Universität des Saarlandes, Verbmobil Report 226, Mai 1997.

E.Amaldi, M.Mattavelli. *A combinatorical 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.

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

B.Buschbeck-Wolf. *Resolution on Demand.* Universität Stuttgart. Verbmobil Report 196. May 1997.

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

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

M.Emele, M.Dorna. *Ambiguity Preserving Machine Translation using Packed Representations.* In Proceedings of the 17th International Conference on Computational Linguistics (COLING-ACL '98), Montreal, Canada. August 1998.

R.Frederking, S.Nirenburg. *Three Heads are Better than One*, ANLP94P, p 95-100, 1994.

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

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

S.Koch, U.Küssner, M.Stede. *Contextual Disambiguation*, In W.Wahlster, Ed. *Verbmobil: Foundations of Speech to Speech Translation* Springer Verlag, 2000.

U.Küssner. *Applying DL in Automatic Dialogue Interpreting*, Proceedings of the International Workshop on Description Logics - DL-97, pp 54-58, Gif sur Yvette, France, 1997.

U.Küssner. *Description Logic Unplugged* Proceedings of the International Workshop on Description Logics - DL-98, pp 142-146, Trento, Italy, 1998.

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

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

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

M.Schiehlen, J.Bos, M.Dorna *Verbmobil Interface Terms (VITs)*, In W.Wahlster, Ed. *Verbmobil: Foundations of Speech to Speech Translation* Springer Verlag, 2000.

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

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

W.Wahlster, Ed. *Verbmobil: Foundations of Speech to Speech Translation* Springer Verlag, 2000.

C.Worm, C.J.Rupp. *Towards Robust Understanding of Speech by Combination of Partial Analyses* Proceedings of ECAI 1998