What Can Machine Translation Learn from Speech Recognition?

Franz Josef Och[1,2]

[1]AIXPLAIN AG
Monnetstr. 18
52146 Würselen, Germany
f.j.och@aixplain.de

Hermann Ney[2]

[2]Lehrstuhl für Informatik VI
Computer Science Department
RWTH Aachen - University of Technology
52056 Aachen, Germany
{och,ney}@informatik.rwth-aachen.de

Abstract

The performance of machine translation technology after 50 years of development leaves much to be desired. There is a high demand for well performing and cheap MT systems for many language pairs and domains, which automatically adapt to rapidly changing terminology. We argue that for successful MT systems it will be crucial to apply data-driven methods, especially statistical machine translation. In addition, it will be very important to establish common test environments. This includes the availability of large parallel training corpora, well defined test corpora and standardized evaluation criteria. Thereby research results can be compared and this will open the possibility for more competition in MT research.

1 Introduction

There is an increasing demand for machine translation systems which produce high quality translations and which can be easily adapted to many language pairs, new domains and changing terminology. During the recent two decades a lot of progress has been made, yet the current quality of MT systems still leaves much to be desired. The development of an MT system for a new language pair or a new domain is very time-consuming and expensive.

We believe that the quality of MT systems suffers from the lack of data-driven methods in the mainstream of machine translation research. We suggest using statistical methods for automatically learning machine translation knowledge. We expect that in a few years this will be the dominant approach to develop machine translation systems. Hence, it is interesting to think about the effects on the development of machine translation systems. In addition, we point out some important research issues that need to be addressed in order to obtain better MT quality. Finally, a main obstacle towards a more efficient MT research community is the lack of competition. Therefore, it is very important to install common test environments in the MT community.

We expect that in machine translation we will see a similar development as in speech recognition about thirty years ago. At that time statistical methods have been introduced in speech recognition systems, which resulted in a tremendous improvement in recognition accuracy in the eighties. Today statistical methods are the mainstream approach in speech recognition. We believe that it is possible to carry over to machine translation research some of the paradigms that revealed to be important in speech recognition research.

In Section 2, we will give an overview of the main advantages of this approach. The architecture and the development cycle of a statistical MT system will be described in Section 3. In Section 4 we will present evaluation results showing the high quality that can be obtained by using statistical methods. Yet, present approaches to statistical machine translation have some limitations that need to be dealt with in order to obtain an additional improvement in translation quality. This will be described in Section 5. A main problem in the MT community is the lack of common training/test corpora and evaluation criteria. Possible solutions to this problem will be presented in Section 6.

2 Statistical Machine Translation

The use of statistics in computational linguistics has been extremely controversial for more than three decades. The controversy is very well summarized by the statement of Chomsky in 1969 (Chomsky, 1969):

> "It must be recognized that the notion of a 'probability of a sentence' is an entirely useless one, under any interpretation of this term".

This statement was considered to be true by the majority of experts from artificial intelligence and computational linguistics, and the concept of statistics was banned from computational linguistics for many years.

What is overlooked in this statement is the fact that in an automatic system for speech recognition or text translation, we are faced with the problem of taking

decisions. It is exactly here where statistical decision theory comes in. In automatic speech recognition (ASR), the success of the statistical approach is based on the equation:

$$\begin{aligned} \text{ASR} &= \text{Acoustic–Linguistic Modeling} \\ &\quad + \text{Statistical Decision Theory} \end{aligned}$$

Similarly, for machine translation, the statistical approach is expressed by the equation:

$$\begin{aligned} \text{MT} &= \text{Linguistic Modeling} \\ &\quad + \text{Statistical Decision Theory} \end{aligned}$$

For the 'low-level' description of speech and image signals, it is widely accepted that the stochastic framework allows an efficient coupling between the observations and the models, which is often described by the buzz word 'subsymbolic processing'. But there is another advantage in using probability distributions in that they offer an explicit formalism for expressing and combining hypothesis scores:

- The probabilities are directly used as scores: These scores are normalized, which is a desirable property: when increasing the score for a certain element in the set of all hypotheses, there must be one or several other elements whose scores are reduced at the same time.
- It is evident how to combine scores: depending on the task, the probabilities are either multiplied or added.
- Weak and vague dependencies can be modeled easily. Especially in spoken and written natural language, there are nuances and shades that require 'grey levels' between 0 and 1.

Even if we think we can manage without statistics, we will need models that always have some free parameters. Then the question is how to train these free parameters. The obvious approach is to adjust these parameters in such a way that we get optimal results in terms of error rates or similar criteria on a representative sample. So we have made a complete cycle and have reached the starting point of the stochastic modeling approach again!

When building an automatic system for speech or language, we should try to use as much prior knowledge as possible about the task under consideration. This knowledge is used to guide the modeling process and to enable improved generalization with respect to unseen data. Therefore, in a good stochastic modeling approach, we try to identify the common patterns underlying the observations, i.e. to capture dependencies between the data in order to avoid the pure 'black box' concept.

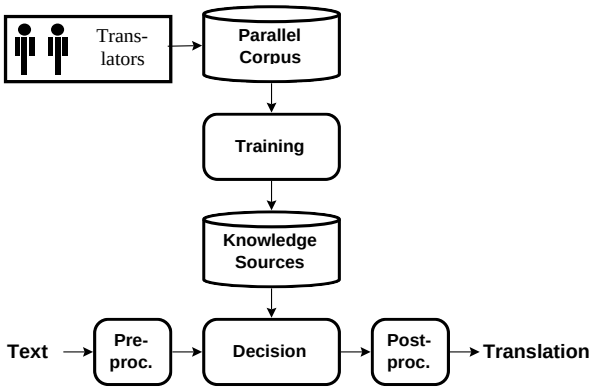

Figure 1: Architecture of a statistical MT system.

3 Rapid Prototyping of Statistical MT Systems

Figure 1 shows the architecture of a statistical machine translation system. A main advantage of the statistical approach to machine translation is the fact that it is possible to develop very quickly new MT systems for new language pairs and new domains under the assumption that a suitable amount of training data is available. In classical rule-based systems, the knowledge sources used in the translation process have to be provided by hand from linguistic experts. In a fully fledged data-driven approach, the starting point is a parallel training corpus which consists of translation examples which were produced by human translators. In the training phase the necessary training sources are trained automatically. The search or decision process has to achieve an optimal combination of the knowledge sources in order to perform an optimal translation. In addition, we may explicitly allow optional transformations (pre-/postprocessing) to simplify the translation task for the algorithm.

Figure 2 presents the development cycle of a statistical MT system. The first step is the collection of training data. The second step is the automatic training of the system. The output of this step is an operative MT system which typically achieves a reasonable translation quality. Normally, this step is quite fast (see e.g. MT in a day experiment (Al-Onaizan et al., 1999)). Afterwards, the system is tested and an error analysis is performed. Depending on the result of this error analysis various modifications are performed:

- **Better pre-/postprocessing**: Various natural language phenomena are notoriously difficult to handle for state-of-the-art statistical technology. One method for dealing with this problem is to preprocess the data such that it is better suited for the statistical translation models. Here classical rule-based MT technology can be used.

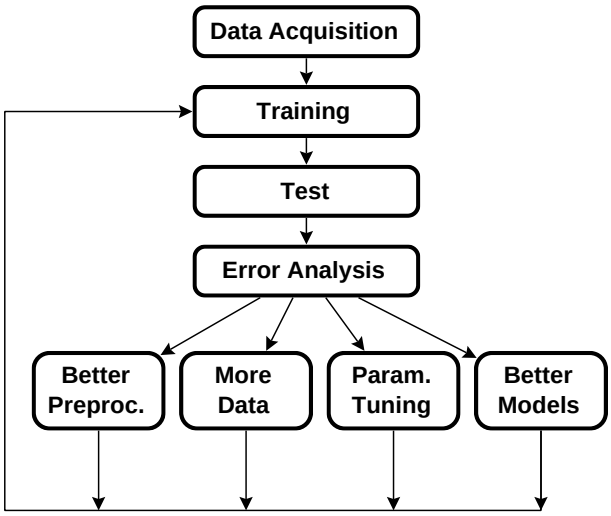

Figure 2: Development cycle of a statistical MT system.

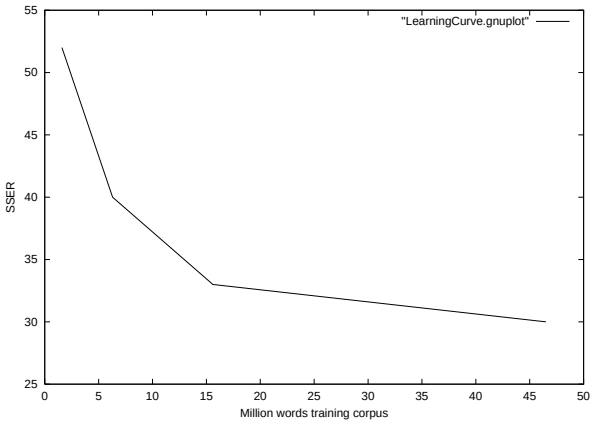

Figure 3: Learning curve of the alignment template system on the Hansards task. The subjective error rate (SSER) has been evaluated using the tool described in (Nießen et al., 2000)

- **More training data**: The learning curve of an MT system shows how much training data is needed to obtain a certain performance. Figure 3 shows an example of a learning curve obtained by the alignment template system (Och et al., 1999) for different amounts of the Hansards corpus. In this system, the error rate improves by about 4 % if the size of the training corpus increases by a factor of two.

- **Tuning of system parameters**: Here, various system parameters such as the relative weight of the translation model vs. the language model can be adjusted so that the error rate is optimized. To do this efficiently, it is important to have the possibility of cheaply evaluating a large number of different translation results. We will deal with this problem in Section 6.

- **Development of better models**: Here lies the art of statistical machine translation: developing models which better capture the properties of natural language and whose free parameters can be estimated reliably from training data.

4 Experimental Results

Whereas stochastic modeling is widely used in speech recognition, there are so far only a few research groups that apply stochastic modeling to language translation (Brown et al., 1993; Berger et al., 1994; Och and Weber, 1998; Alshawi et al., 1998; Wang and Waibel, 1998; Knight, 1999; Ney et al., 2000a). The presentation here is based on work carried out in the framework of the EUTRANS project (Casacuberta et al., 2001) and the VERBMOBIL project (Wahlster, 2000). The goal of the VERBMOBIL project is the translation of spoken dialogues in the domains of appointment scheduling and travel planning. The languages are German and English.

Whereas during the progress of the project many offline tests were carried out for the optimization and tuning of the system, the most important evaluation was the final evaluation of the VERBMOBIL prototype in spring 2000. This end-to-end evaluation of the VERBMOBIL system was performed at the University of Hamburg (Tessiore and v. Hahn, 2000). In each session of this evaluation, two native speakers conducted a dialogue. They did not have any direct contact and could only interact by speaking and listening to the VERBMOBIL system.

In addition to the statistical approach, three other translation approaches had been integrated into the VERBMOBIL prototype system (Wahlster, 2000):

- a classical transfer approach,
  which is based on a manually designed analysis grammar, a set of transfer rules, and a generation grammar,

- a dialogue act based approach,
  which amounts to a sort of slot filling by classifying each sentence into one out of a small number of possible sentence patterns and filling in the slot values,

- an example based approach,
  where a sort of nearest neighbor concept is applied to the set of bilingual training sentence pairs after suitable preprocessing.

In the final end-to-end evaluation, human evaluators judged the translation quality for each of the four

Table 1: Error rates of spoken sentence translation in the VERBMOBIL end-to-end evaluation.

Translation Method	Error [%]
Semantic Transfer	62
Dialogue Act Based	60
Example Based	52
Statistical	29

translation results using the following criterion:

Is the sentence approximately correct: yes/no?

The evaluators were asked to pay particular attention to the semantic information (e.g. date and place of meeting, participants etc) contained in the translation. A missing translation as it may happen for the transfer approach or other approaches was counted as wrong translation. The evaluation was based on 5069 dialogue turns for the translation from German to English and on 4136 dialogue turns for the translation from English to German. The speech recognizers used had a word error rate of about 25%. The overall sentence error rates, i.e. resulting from recognition *and* translation, are summarized in Table 1. As we can see, the error rates for the statistical approach are smaller by a factor of about 2 in comparison to the other approaches.

5 Towards Better Systems

More sophisticated statistical models

There are some obvious extensions to state-of-the-art statistical MT systems:

- There is a need for more sophisticated models which are better suited for the recursive structure of natural languages. Current statistical machine translation systems have problems with nonlocal phenomena, i.e. dependencies between nonconsecutive words and there are only a few approaches that try to deal with this problem (Wu, 1994; Alshawi et al., 1998; Wang and Waibel, 1998).

- Statistical translation systems typically ignore the context in which a sentence appears. This means that e.g. anaphora are normally translated by their most probable translation, which is a source of systematic errors. In addition, most models do not depend on the text structure or dialogue-act (in speech translation).

- The use of morphological processing should be part of the translation process (Nießen and Ney, 2000). A problem of morphologically rich languages is that the statistical approach has a significant problem with sparse data if morphology is not handled.

It should be emphasized that e.g. the Verbmobil system results presented in the previous section were obtained with a system that suffers to a certain extent from all these problems of state-of-the-art SMT technology. Therefore, from a fully fledged statistical machine translation system we expect an additional significant gain in translation quality.

Automated collection of training data

A key element in the data-driven approach to machine translation is the collection of large amounts of useful training data. A very interesting approach is the idea of automatically collecting large amounts of training data from the Internet (Resnik, 1999). One of the problems of the data collected in this way is that it frequently contains wrong translations, omissions and other noise. In order to make use of this data, it is important to apply robust sentence alignment, automatic detection and filtering of wrong translation examples.

Fine-grained combination of the statistical and the linguistic approach

There is hardly any system that performs a fine-grained combination of statistical and rule-based systems. There are systems that try to improve translation quality by combining different MT systems by an independent translation with different systems and deciding afterwards which translation is to be used (Nirenburg and Frederking, 1994; Cavar et al., 2000). Yet, the improvements obtained by these approaches are typically small.

A more significant improvement can be expected by an incorporation of linguistic knowledge sources into statistical models. This means that in addition to the bilingual corpus used to train the translation model additional knowledge sources (e.g. parallel tree banks, WordNet, ...) will be used in the training of refined translation models.

6 Towards More Competition

Standard Training-/Test corpora

An important problem in the machine translation community is the lack of a number of suitable training/test corpora, which can be used by various research groups. This is different to the situation in the speech recognition community where it is generally accepted that new approaches have to be evaluated on common training/test corpora. There are various advantages of this approach. It is possible to compare results of different groups and to decide which methods work well and which methods do not. In addition,

there is a competition among various groups to produce better results.

More efforts, e.g. supported by government agencies, are needed to produce such corpora and to make these corpora freely available to interested research groups.

Common low-dimensional evaluation criteria

In order to compare results of different research groups, it is important to use not only common training/test corpora, but also common evaluation criteria. In speech recognition, the word error rate (WER) is the generally accepted evaluation criterion. In machine translation, there does not exist a generally accepted evaluation criterion. Yet, often it is even doubted that it is possible to really quantify translation quality.

Machine translation quality can be measured in a large number of non-orthogonal dimensions (Hovy, 1999). Yet, ideally we would like to have a one-dimensional evaluation criterion as in speech recognition, which makes comparing different systems easy. In addition, we would like to use an evaluation criterion that is cheap in its application.

A general problem of subjective MT evaluation is that the comparability of different results is hard to guarantee if the evaluation is not performed by the same group of humans in the same moment. One method to deal with this problem is the use of common evaluation tools and databases (Jones and Rusk, 2000; Nießen et al., 2000; Vogel et al., 2000). Another possibility is the establishment of a central test agency, which performs the evaluation for various research groups.

There have been some suggestions for automatic evaluation criteria in machine translation such as normal word error rate, position-independent word error rate, multi-reference word error rate. Interestingly, these very simple evaluation criteria often, but not always, correlate with a subjectively evaluated translation quality (Ney et al., 2000a). It seems that these criteria are well suited for comparing different versions of the same system, but cannot be used when completely different systems are compared.

Therefore, in the development cycle of an MT system, one of these objective criteria can usually be used and only from time to time, the expensive subjective evaluation is performed. This approach was successfully applied in the development of the Verbmobil statistical machine translation system (Och et al., 1999; Ney et al., 2000b).

We do not expect that it is possible to develop one evaluation criterion suited for all possible applications of machine translation. For example, for speech translation the evaluation criterion must focus on *understandability*, in interactive machine translation on *post-editing effort* and in fully automatic machine translation *syntax and semantic* are important.

Every evaluation criterion for MT will have some disadvantages because some aspects are not covered perfectly. The same is true in ASR where the word error rate for example does not distinguish between important and non-important words. Yet, it is better to use this evaluation criterion with small shortcomings instead of using many different incomparable evaluation criteria.

Towards MTC: The Machine Translation Conference

As we pointed out, there is a high demand for an increased international collaboration and competition in machine translation research. A very productive way of stimulating more competition in research seems to be in a way like the MUC (Message Understanding Conference) or TREC (Text Retrieval Conference) where various tasks are defined. A central evaluation agency should organize and perform the system evaluation.

A corresponding 'conference' in machine translation, the MTC - Machine Translation Conference - must include well defined standard training-/test environments and a common evaluation methodology. We suggest having various tasks for components of MT systems such as sentence alignment, word alignment, language modeling and word sense disambiguation. In addition, there would be projects which evaluate the translation quality of full MT systems.

7 Conclusions

We have argued that for future successful MT systems statistical machine translation will be a much more important approach than it is today. We have presented recent evaluation results in the Verbmobil project, which show the high level of quality that has been achieved in SMT.

In order to obtain better translation systems we believe that it is not only important to develop better translation models but also to establish common test environments which allow different research groups to compare research results. As specific step towards more competition in the machine translation research community, we suggest to establish the Machine Translation Conference (MTC) similar to the TREC and MUC conferences.

References

Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, D. Purdy, F. J. Och, N. A. Smith, and D. Yarowsky. 1999. Statistical machine translation, final report, JHU workshop.

http://www.clsp.jhu.edu/ws99/projects/mt/final_report/mt-final-report.ps.

H. Alshawi, S. Bangalore, and S. Douglas. 1998. Automatic acquisition of hierarchical transduction models for machine translation. In *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th Int. Conf. on Computational Linguistics*, volume 1, pages 41–47, Montreal, Quebec, Canada, August.

A. Berger, P. Brown, S. D. Pietra, V. D. Pietra, J. Gillett, J. Lafferty, H. Printz, and L. Ures. 1994. The candide system for machine translation. In *Proceedings ARPA Workshop on Human Language Technology*, pages 157–162, Plainsboro, NJ, March.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

F. Casacuberta, D. Llorenz, C. Martinez, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Pico, A. Sanchis, E. Vidal, and J. Vilar. 2001. Speech-to-speech translation based on finite-state transducers. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, May.

D. Cavar, U. Küssner, and D. Tidhar. 2000. From off–line evaluation to on–line selection. In (Wahlster, 2000), pages 599–612.

N. Chomsky. 1969. Quine's empirical assumptions. In D. Davidson and J. Hintikka, editors, *Words and objections. Essays on the work of W. V. Quine*. Reidel, Dordrecht, The Netherlands.

E. Hovy. 1999. Toward finely differentiated evaluation metrics for machine translation. In *Proceedings of the EAGLES Workshop on Standards and Evaluation*, pages 127–133, Pisa, Italy.

D. A. Jones and G. M. Rusk. 2000. Toward a scoring function for quality-driven machine translation. In *COLING '00: The 18th Int. Conf. on Computational Linguistics*, pages 376–382, Saarbrücken, Germany, August.

K. Knight. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4):607–615.

H. Ney, S. Nießen, F. J. Och, H. Sawaf, C. Tillmann, and S. Vogel. 2000a. Algorithms for statistical translation of spoken language. *IEEE Trans. on Speech and Audio Processing*, 8(1):24–36, January.

H. Ney, F. J. Och, and S. Vogel. 2000b. Statistical translation of spoken dialogues in the verbmobil system. In *Workshop on Multi-Lingual Speech Communication*, pages 69–74, Kyoto, Japan, October.

S. Nießen and H. Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *COLING '00: The 18th Int. Conf. on Computational Linguistics*, pages 1081–1085, Saarbrücken, Germany, July.

S. Nießen, F. J. Och, G. Leusch, and H. Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, pages 39–45, Athens, Greece, May.

S. Nirenburg and R. Frederking. 1994. Toward multi-engine machine translation. In *Proceedings ARPA Workshop on Human Language Technology*, pages 147–151, Plainsboro, New Jersey, March.

F. J. Och and H. Weber. 1998. Improving statistical natural language translation with categories and rules. In *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th Int. Conf. on Computational Linguistics*, pages 985–989, Montreal, Canada, August.

F. J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

P. Resnik. 1999. Mining the web for bilingual text. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 527–534, University of Maryland, College Park, MD, June.

L. Tessiore and W. v. Hahn. 2000. Functional validation of a machine interpretation system: Verbmobil. In (Wahlster, 2000), pages 611–631.

S. Vogel, S. Nießen, and H. Ney. 2000. Automatic extrapolation of human assessment of translation quality. In *2nd International Conference on Language Resources and Evaluation (LREC 2000): Proceedings of the Workshop on Evaluation of Machine Translation, pp*, pages 35–39, Athens, Greece, May/June.

W. Wahlster, editor. 2000. *Verbmobil: Foundations of speech-to-speech translations*. Springer-Verlag, Berlin.

Y.-Y. Wang and A. Waibel. 1998. Modeling with structures in statistcal machine translation. In *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th Int. Conf. on Computational Linguistics*, volume 2, pages 1357–1363, Montreal, Quebec, Canada, June.

D. Wu. 1994. Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proc. of the 32nd Annual Conf. of the Association for Computational Linguistics*, pages 80–87, June.