One Decade of Statistical Machine Translation: 1996–2005

Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
ney@informatik.rwth-aachen.de

Abstract

In the last decade, the statistical approach has found widespread use in machine translation both for written and spoken language and has had a major impact on the translation accuracy. This paper will cover the principles of statistical machine translation and summarize the progress made so far.

1 Introduction

The goal of this paper is to cover the state of the art in statistical machine translation (MT). We will re-visit the underlying principles of the statistical approach to machine translation (and other tasks in natural language processing) and summarize the progress that has been made over the last decade.

2 Recasting MT as a Statistical Problem

2.1 The Baseline Approach to Statistical MT

When translating a sentence from a source language into a target language, a human translator has to take into account the following considerations:

- Lexical choice: Typically, a source word can have several translations in the target language, and the selection of the suitable translation depends on the context.
- Position re-ordering: The word positions in source and target sentence are different.
- Syntax and semantics: To generate the target sentence, syntactic and semantic constraints have to be taken into account.

The corresponding operations are summarized in Fig. 1. The traditional non-statistical approach to MT is to manually design rules and knowledge sources for these operations. There are two open problems with this concept: 1. How can we get hold of all the rules and make sure that we know all the rules that the system needs? 2. How can we achieve a coherent and consistent interaction of all these rules when generating the target sentence?

Both problems are elegantly addressed in the statistical approach. Instead of using hard rules, we make use of *probability distributions* that serve as probabilistic knowledge sources. In the baseline version of the statistical approach as introduced by IBM (Berger et al. 1994; Brown et al. 1993), the translation task can be expressed as follows. We are given a sentence F in the source language and we want to generate the corresponding target sentence E. For this purpose, we consider the associated posterior distribution $p(F|E)$ of all possible pairs (F, E) and select the target sentence $\hat{E}(F)$ with the highest posterior probability:

$$\begin{aligned} F \rightarrow \hat{E}(F) &= \arg\max_E \{p(E|F)\} \\ &= \arg\max_E \{p(E) \cdot p(F|E)\} \end{aligned}$$

which is referred to as Bayes decision rule. In the second equation, the posterior probability $p(F|E)$ has been replaced by the joint probability $p(E, F)$ which is written as the product of the so-called *language model* $p(E)$ and the so-called *translation model* $p(F|E)$.

The translation model is still rather complex because it describes probability distributions over *whole sentences* F and E. To reduce its complexity, we make use of the concept of *word alignments* A that capture the correspondences between source and target words and allow us to move from the sentence level to the word level. Since these correspondences are not unique or deterministic, they are described by probability distributions as well and we obtain for the translation probability $p(F|E)$:

$$p(F|E) = \sum_A p(A|F) \cdot p(F|E, A)$$

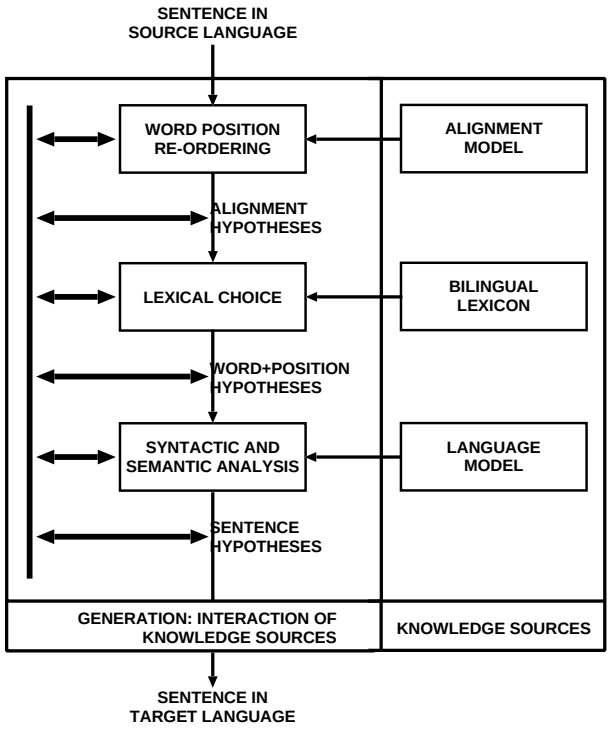

Figure 1: Architecture of a translation system.

with the alignment model $p(A|F)$ and the lexicon model $p(F|E,A)$. In an influential paper, the IBM group (Brown et al. 1993) designed a series of five alignment-lexicon models of increasing complexity that today still serve as the starting point for virtually all statistical MT systems.

The typical advantage in using probability distributions is that they offer an explicit formalism for expressing and combining hypothesis scores:

- The probabilities can be 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 dependences can be modelled easily. Particularly in spoken and written natural language, there are nuances and shades that require 'grey levels' between 0 and 1.
- There are powerful algorithms for learning these probabilities automatically without human intervention (see later).

To solve a practical problem like MT (or other tasks in natural language processing), we think that the statistical approach cannot be avoided on principal grounds. Suppose we have designed a *non-statistical* approach, i.e. an approach that explicitly avoids probability distributions. This approach will have a certain number of 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. But this is exactly the starting point for the statistical approach along with Bayes decision rule!

When building a statistical system, we will try to use as much prior knowledge as possible. The concept of *alignments* is an example of such prior knowledge, i.e. we know that the correspondence of source-target sentence pairs can be reduced to correspondences at the level of words or word groups. The prior knowledge will help us to design suitable probabilistic models and to improve the generalization with respect to unseen data. Therefore in a good statistical approach, we will 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.

2.2 The Tasks in Statistical MT

From a general point of view, there are four tasks to be addressed in the statistical approach as illustrated in Fig. 2:

- **error measure and decision rule:**
  Bayes decision rule is based on minimizing the so-called posterior risk which requires a quantitative error measure or cost function. In statistical MT (as in speech recognition), the traditional error measure is the 0/1-loss function that minimizes the *sentence* error rate but not necessarily the error measures used in translation like WER, PER and BLEU (see later). This type of inconsistency is not addressed in the literature with the exception of (Kumar and Byrne 2004). Furthermore, some applications of machine translation require criteria at the word level (like confidence

measures (Ueffing et al. 2003)) rather than at the sentence level.

- **probability models:**
  The probability models are used to replace the true but unknown probability distributions in Bayes decision rule. Their ultimate goal is to provide the link between the input data (source sentence) and the output data (target sentence) that have to be produced by the translation system. It is exactly here where linguistic knowledge will be helpful to come up with better models in the future.
- **training criterion:**
  The training criterion is used to learn the free parameters of the probability models from the training data. As in speech recognition, the popular training criterion is the maximum likelihood criterion, often in an EM-type algorithm. Improvements can be expected for training criteria that are more discriminative and that better match the translation specific error measures.
- **decision rule:**
  According to the type of Bayes decision rule used, the generation (or search) algorithm selects the most suitable target sentence from the huge number of all possible target sentences. Due to the combinatorial complexity of this task, an efficient implementation is crucial. Again, as in speech recognition, all generation (or search) algorithms for translation have used the maximum sentence probability as search criterion.

To summarize, the statistical approach defines a framework within which we still have to work out the details of the decision rule, the probability models, the training criterion and the generation algorithm.

3 Statistical MT in 1996–2005

3.1 Projects

The principles on which today's statistical systems for machine translation are based were worked out only around 1990 and described in (Brown et al. 1993). Considerable progress has been made since then in both *written* and *spoken* language translation.
Spoken language translation has been and is being investigated in a number of joint projects

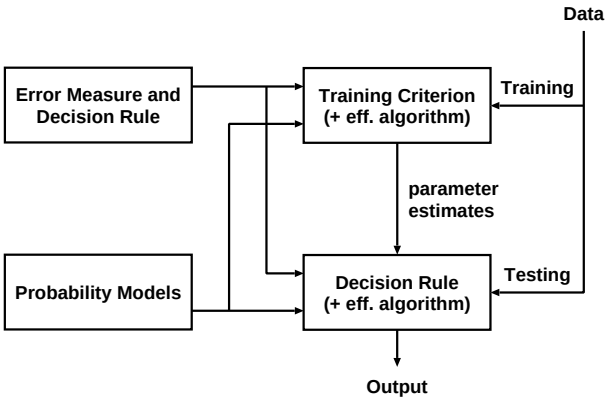

Figure 2: Tasks in statistical MT.

at some national levels, the European level and the international level (*C-Star, ATR, Verbmobil, Eutrans, LC-Star, PF-Star, TC-Star, ...*). Apart from the project TC-Star, which started only recently, all these projects addressed translation tasks with rather limited domains (like traveling and tourism) and medium-sized vocabularies (about 10 000 words). The best performing translation systems are based on various types of statistical approaches (Och and Ney 2002) including example-based methods (Sumita et al. 2003), finite-state transducers (Casacuberta and Vidal 2004) and other data driven approaches. This is the characteristic and most striking result of the various projects.
A similar experience was made for *written language translation.* Within the US Tides program, the goal of the MT project was to translate news articles from Chinese to English and from Arabic to English, which implied large vocabularies (80 000 and more words) and rather unrestricted domains. In the evaluations, it was found that, due to the recent improvements, the statistical approach was able to produce competitive or superior results in comparison with conventional translation systems. This is remarkable because the statistical systems for these language pairs were developed within a short period of only a few years.

3.2 Major Improvements

The progress achieved over the last 10 years is due to several factors:

- **automatic error measures:**
  At the time the IBM group introduced the statistical approach to MT, it was not clear how to automatically evaluate the

quality of MT output. Meanwhile, various measures have been suggested and tested experimentally, like BLEU, WER (word error rate) and PER (position independent word error rate) (Papineni et al. 2003; Tillmann et al. 1997). Although it is well known that none of these measures are perfect, they seem to correlate well with MT quality when measured in terms of adequacy and fluency. At the present level of MT performance, most researchers consider them to be adequate to assess the progress in the field. Such automatic measures are very important to allow for fast train/test cycles in research.

- **efficient algorithms for training:**
  The alignment-lexicon models are trained on large sets of source-target sentence pairs. Although its principle was already introduced in 1993 (Brown et al. 1993), the training procedure is fairly complex and its details were studied experimentally only later (Och and Ney 2003). This work resulted in a public software package GIZA++ that is used by most researchers in the field. Furthermore, a couple of refinements were introduced beyond the original alignment-lexicon models like the so-called HMM approach (Vogel et al. 1996). Another refinement is the symmetrization of the training procedure, where the role of target and source sentences is exchanged in order to improve the quality of the final word alignments.

- **context dependent or phrase-based lexicon models:**
  The original alignment-lexicon models (Brown et al. 1993) do not take into account the context in which both the source and the target words appear. There is an evident need to introduce more context dependencies into these models, e.g. by handling word groups and phrases rather than single words. There have been a number of successful extensions that move away from single words and handle word groups in both the source and target language (Och et al. 1999; Zens et al. 2002; Koehn et al. 2003; Och and Ney 2004). Typically, these extensions seem to be limited to the extraction of bilingual phrases *after* the alignment-lexicon models (like IBM and HMM) have been trained. In other words, the phrase-based models are not yet incorporated into the iterative training procedure. It is interesting to note that this extraction of phrase pairs shows a certain similarity to *memory-based* translation. The important difference, however, is that in statistical MT these pairs are extracted *automatically*.

- **efficient algorithms for generation:**
  To generate the target sentence, various strategies have been studied like A^* search and dynamic programming beam search (Koehn 2004; Tillmann and Ney 2003). In the experimental tests, dynamic programming beam search has been found to be much more efficient than a (pure) A^* strategy. Typically, this beam search strategy works by processing the source positions in a left-to-right fashion and allowing word re-orderings to a certain degree. Furthermore, these search strategies have been extended to produce word lattices and N-best lists rather than only the single best sentence.

- **log-linear model combinations and re-scoring:**
  The baseline models in statistical translation are the lexicon model, the alignment model and the language model. Due to model and training shortcomings, it is convenient to assign 'relevance' factors (or scaling exponents) to these models by combining them in a log-linear fashion. In conjunction with N-best lists, these scaling exponents can be trained automatically (Och and Ney 2002). Furthermore, in this framework, additional types of dependencies (referred to as *feature functions)* can be integrated easily.

- **more powerful computers and more parallel corpora:**
  Both the training and the generation algorithms in statistical MT require a huge computing power, in particular the memory requirements tend to be very demanding. Furthermore, the bilingual corpora have been steadily growing in size.

3.3 Additional Research Directions

Syntax-based Translation Models. While the alignment approach introduced in (Brown et al. 1993) does not make use of any syntactic concepts, there were attempts at introducing

explicit syntactic structures into the statistical translation models (Alshawi et al. 2000; Wu 1997; Yamada and Knight 2001). In such a way, it is expected that the difference in the word order between target and source sentences can be better taken into account. At present, these syntax-based extensions do not (yet) seem to pay off in terms of performance. In addition, the syntactic approach could also include a morphological analysis (Nießen and Ney 2004) so that the statistical approach could go beyond the usual full forms of words.

Speech Translation. The translation of *spoken* language requires the combination of two operations, namely the *recognition* of the spoken source sentence and its *translation* into the target sentence. Thus, we are faced with the additional problem of finding a suitable integration of recognition and translation. While the principles of this integration are more or less well known (Ney 1999), in practice most systems only make use of N-best lists, which however do not seem to improve performance.

Interactive MT. What we have considered so far is called *fully automatic* MT. In *interactive MT*, the user interacts with the system when generating the target sentence. It turns out that Bayes decision rule can be extended to efficiently handle this interactive MT (Och et al. 2003). This was the topic of the European project *TransType 2*.

4 Conclusion

The past experiences with speech and language processing have shown that a substantial amount of progress was always achieved by the improvement of the more or less purely algorithmic concepts of how we model the dependencies of the data and how the system better learns from the data. We expect that future work along these lines will result in significant improvements for statistical MT. So far, the statistical approach to MT has exploited only a small number of the research directions explored for speech recognition.

Acknowledgment. The work reported is based on projects that were supported by the German BMBF (*Verbmobil)*, by the German Science Foundation (DFG) and by various European projects (*Eutrans, LC-Star, PF-Star and TC-Star*).

References

H. Alshawi, S. Bangalore, S. Douglas: Learning Dependency Translation Models as Collection of Finite-State Head Transducers. *Computational Linguistics*, Vol. 26, No. 1, pp. 45–60, 2000.

A. L. Berger, P. F. Brown, J. Cocke et al.: The Candide System for Machine Translation. *ARPA Human Language Technology Workshop*, Plainsboro, NJ, pp. 152-157, March 1994.

P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, R. L. Mercer: The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

F. Casacuberta, E. Vidal: Machine Translation with Inferred Stochastic Finite-State Transducers. *Computational Linguistics*, Vol. 30, No. 2, pp. 205–225, June 2004.

P. Koehn: Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. *Meeting of the American Association for Machine Translation (AMTA)*, Washington DC, pp. 115-124, Sep./Oct. 2004.

P. Koehn, F. J. Och, D. Marcu: Statistical Phrase-Based Translation. *Human Language Technology Conference 2003 (HLT-NAACL 2003)*, Edmonton, Canada, pp. 127-133, May 2003.

S. Kumar, W. Byrne: Minimum Bayes-Risk Decoding for Statistical Machine Translation. *Human Language Technology Conference (HLT-NAACL)*, Boston, MA, pp. 169–176, May 2004.

H. Ney: Speech Translation: Coupling of Recognition and Translation. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* Phoenix, AR, pp. I-517-520, March 1999.

S. Nießen, H. Ney: Statistical Machine Translation with Scarce Resources using Morpho-syntactic Information. *Computational Linguistics*, Vol. 30, No. 2, pp. 181–204, June 2004.

F. J. Och, H. Ney: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. *Annual Meeting of the Ass. for Computational Linguistics (ACL)*, Philadelphia, PA, pp. 295-302, July 2002.

F. J. Och, H. Ney: A Systematic Comparison of Various Alignment Models. *Computational*

Linguistics, Vol. 29, No. 1, pp. 19-51, March 2003.

F. J. Och, H. Ney: The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, Vol. 30, No. 4, pp. 417-449, Dec. 2004.

F. J. Och, C. Tillmann, H. Ney: Improved Alignment Models for Statistical Machine Translation. *Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD, pp. 20-28, June 1999.

F. J. Och, R. Zens, H. Ney: Efficient Search for Interactive Statistical Machine Translation. *Europ. Assoc. for Comput. Linguistics (EACL)*, Budapest, Hungary, pp. 387-393, April 2003.

K. Papineni, S. Roukos, T. Ward, W.-J. Zhu: BLEU: a Method for Automatic Evaluation of Machine Translation. *40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, pp. 311–318, July 2002.

E. Sumita, Y. Akiba, T. Doi et al.: A Corpus-Centered Approach to Spoken Language Translation. *Conf. of the Europ. Chapter of the Ass. for Computational Linguistics (EACL)*, Budapest, Hungary, pp. 171–174 (Conference Companion), April 2003.

C. Tillmann, H. Ney: Word Re-Ordering and a DP Beam Search Algorithm for Statistical Machine Translation. *Computational Linguistics*, Vol. 29, No. 1, pp. 97–133, March 2003.

C. Tillmann, S. Vogel, H. Ney et al.: Accelerated DP-based Search for Statistical Translation. *Fifth European Conf. on Speech Communication and Technology,* Rhodos, Vol. 5, pp. 2667-2670, Sep. 1997.

N. Ueffing, K. Macherey, H. Ney: Confidence Measures for Statistical Machine Translation. *Machine Translation Summit IX*, New Orleans, LO, pp. 394–401, Sep. 2003.

S. Vogel, H. Ney, C. Tillmann: HMM-Based Word Alignment in Statistical Translation. *Int. Conf. on Computational Linguistics*, Copenhague, Denmark, pp. 836-841, Aug. 1996.

D. Wu: Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, Vol. 23, No. 3, pp. 377–403, 1997.

K. Yamada, K. Knight: A Syntax-based Statistical Translation Model. *Annual Meeting of the Ass. for Computational Linguistics*, Toulouse, France, pp. 523–530, July 2001.

R. Zens, F. J. Och, H. Ney: Phrase-Based Statistical Machine Translation. *German Conf. on Artificial Intelligence (KI 2002),* Aachen, Germany, Springer, LNAI, pp. 18-32, Sep. 2002.