The Statistical Approach to Machine Translation and a Roadmap for Speech Translation

Hermann Ney

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

Abstract

During the last few years, the statistical approach has found widespread use in machine translation, in particular for spoken language. In many comparative evaluations of automatic speech translation, the statistical approach was found to be significantly superior to the existing conventional approaches. The paper will present the main components of a statistical machine translation system (such as alignment and lexicon models, training procedure, generation of the target sentence) and summarize the progress made so far. We will conclude with a roadmap for future research on spoken language translation.

1. Introduction

The automatic translation of language is generally referred to as *machine translation*. Typically, this term is used for *written language* or *text* input, where the implicit assumption is that the input is uncorrupted, i.e. without errors. This task is very much different from *spoken speech* input, where the system must cope with speech recognition errors and also the ungrammatical structure of spoken language.

The translation of *spontaneous speech* poses additional difficulties for the task of automatic translation. Typically, these difficulties are caused by errors of the recognition process, which is carried out before the translation process. As a result, the sentence to be translated is not necessarily well-formed from a syntactic point-of-view. Even without recognition errors, speech translation has to cope with a lack of conventional syntactic structures because the structures of spontaneous speech differ from those of written language.

The statistical approach shows the potential to tackle these problems for various reasons. First, the statistical approach is able to avoid hard decisions at any level of the translation process. Second, for any source sentence, a translated sentence in the target language is guaranteed to be generated. This will be hopefully a syntactically correct sentence in the target language; but even if this is not the case, the translated sentence will often convey the meaning of the spoken sentence.

2. State of the Art

Until today, spoken language translation has been investigated in a number of joint projects at some national levels, the European level and the international level (C-Star, ATR, Verbmobil, Eutrans, Nespole!, Fame, LC-Star, PF-Star, ...). These systems are still limited in many ways [15, 18]:

- They are able to handle only restricted domains (like appointment, conference registration, travelling and/or tourism information).
- The vocabulary is restricted to about 5 000 to 10 000 words.
- Even for the best performing systems and approaches, fairly high sentence error rates are reported [7, 16].

The best performing translation systems are based on various types of statistical approaches [7, 16] including example-based methods [14], finite-state transducers [4] and other data driven approaches. This is the characteristic and most striking result of the various projects.

The principles on which the statistical approach is based were worked out only around 1990 [3]. Considerable progress has been made since then due to improvements in the underlying models and algorithms and to the availability of bilingual parallel corpora and greater processing power. Recently, for *written* language translation with large vocabularies (about 50 000 words), it was also found that, as a result of this progress, the statistical approach is able to produce competitive results with conventional translation systems that had been optimized over decades, e.g. for Chinese-English translation [9].

3. Statistical Approach

3.1. Principle

From a statistical point of view, the goal of a translation system is to generate the most probable target sentence given the source sentence, i.e. to maximize the posterior probability $Pr(e_1^I|f_1^J)$ over all possible target sentences $e_1^I = e_1...e_i...e_I$ ('English') with length I for the given source sentence $f_1^J = f_1...f_i...f_J$ ('French'). Typically, this posterior probability is re-written as the product of the language model $Pr(e_1^I)$ in the target language and the translation model $Pr(f_1^J|e_1^I)$:

$$\begin{aligned}(\hat{I}, \hat{e}_1^{\hat{I}}) &= \arg\max_{I,e_1^I} \{Pr(e_1^I|f_1^J)\} \\ &= \arg\max_{I,e_1^I} \{Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)\}\end{aligned}$$

From this point of view, we highlight the main research challenges for spoken language translation:

- *translation model*: Find suitable structures for the translation model $Pr(f_1^J|e_1^I)$, which may be decomposed into the so-called lexicon model that describes the probabilistic relationships between source and target words, and the so-called alignment model that describes the probabilistic relationships between the positions of the words in the source language and of the words in the target language.
- *language model:* Find suitable structures for $Pr(e_1^I)$, i.e. model the redundancy of the target sentence e_1^I at all levels (lexical, syntactic, semantic,...).
- *generation (or search) task*: The target sentence e_1^I with the maximum probability has to be determined. This generation process must be able to handle large vocabularies.
- *statistical learning*: The parameters of the language, alignment and lexicon models must be learned from example data and resources such as monolingual and bilingual training data, bilingual dictionaries, morpho-syntactic analysers etc.
- *integration of recognition and translation*: When handling speech input rather than text input, the statistical approach has to take into account both the ambiguities of the speech recognition process and the disfluencies of the spoken language.

The overall architecture of the statistical translation approach is summarized in Fig. 1. There may be optional transformations to make the translation task simpler for the algorithm. The transformations may range from the categorization of single words and word groups to more complex preprocessing steps that require some parsing of the source string.

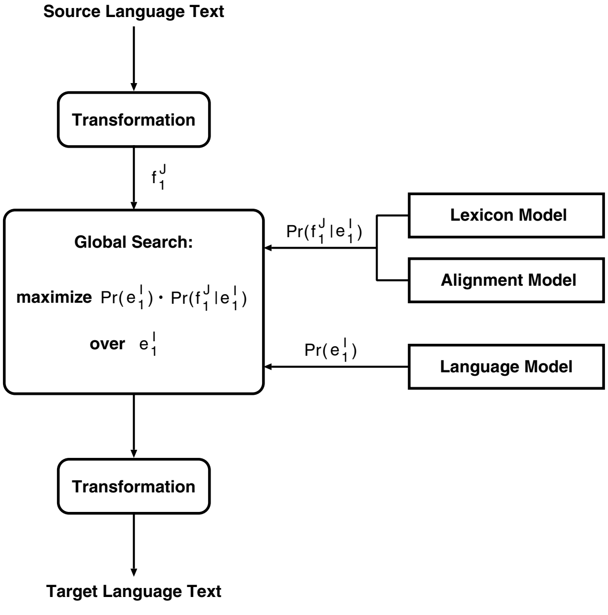

Figure 1: Bayes decision rule for translation.

3.2. Related Approaches

There are a number of related approaches that are also corpus-based and therefore closely related to the statistical approach:

- finite-state approaches [1, 4]:
  Here, the probabilistic dependences are represented by finite-state structures that can be learned automatically from training data.
- example-based approaches [13, 14]:
  In example-based approaches, large bilingual chunks are excised from the set of bilingual sentence pairs. In the translation process, the most similar chunk in the set of source-language chunks is determined, and its corresponding target-language chunk is used as translation. This baseline variant may be refined in various ways to introduce generalization capabilities.
- syntax-based statistical approaches [1, 20, 21]:
  These approaches are obtained as an extension of the statistical approach, where syntactic structures are incorporated into the baseline statistical approach, in particular into the alignment models. The syntactic structure may be modelled in the target language only or in both target and source language.

4. Alignment and Lexicon Models

4.1. Concept

A key issue in modelling the string translation probability $Pr(f_1^J|e_1^I)$ is the question of how we define the correspondence between the words of the target sentence and the words of the source sentence. In typical cases, we can assume a sort of pairwise dependence by considering all word pairs (f_j, e_i) for a given sentence pair $(f_1^J; e_1^I)$. A family of such *alignment models* (IBM-1,...,IBM-5) was developed in [3]. Using the similar principles as in Hidden Markov models (HMM) for speech recognition, we re-write the translation probability by introducing the *hidden alignments* $\mathcal{A}$ for each sentence pair $(f_1^J; e_1^I)$:

$$Pr(f_1^J|e_1^I) \quad = \quad \sum_{\mathcal{A}} Pr(f_1^J, \mathcal{A}|e_1^I)$$

Hidden Markov Models. The first type of alignment models is virtually identical to HMMs and is based on a mapping $j \rightarrow i = a_j$, which assigns a source position j to a target position $i = a_j$. Using suitable modelling assumptions [3, 10], we can decompose the probability $Pr(f_1^J, \mathcal{A}|e_1^I)$ with $\mathcal{A} = a_1^J$:

$$Pr(f_1^J, a_1^J|e_1^I) = p(J|I) \cdot \prod_{j=1}^{J} \left[p(a_j|a_{j-1}, I, J) \cdot p(f_j|e_{a_j})\right]$$

with the length model $p(J|I)$, the alignment model $p(i|i', I, J)$ and the lexicon model $p(f_j|e_i)$. The alignment models IBM-1 and IBM-2 are obtained in a similar way by allowing only zero-order dependencies.

Inverted Alignment Models. For the generation of the target sentence, it is more appropriate to use the concept of *inverted alignments* which perform a mapping from a target position i to a *set* of source positions j, i.e. we consider mappings B of the form:

$$B : i \rightarrow B_i \subset \{1, ..., j, ..., J\}$$

with the constraint that each source position j is covered exactly once. Using such an alignment $\mathcal{A} = B_1^I$, we re-write the probability $Pr(f_1^J, \mathcal{A}|e_1^I)$:

$$Pr(f_1^J, B_1^I|e_1^I) = p(J|I) \cdot \prod_{i=1}^{I} \Big[p(B_i|B_1^{i-1}) \cdot \prod_{j \in B_i} p(f_j|e_i) \Big]$$

By making suitable assumptions, in particular first-order dependencies for the inverted alignment model $p(B_i|B_1^{i-1})$, we arrive at what is more or less equivalent to the alignment models IBM-3, 4 and 5 [10].

4.2. Training

The unknown parameters of the alignment and lexicon models are estimated from a corpus of bilingual sentence pairs. The training criterion is the maximum likelihood criterion. As usual, the training algorithms can guarantee only local convergence. In order to mitigate the problems with poor local optima, we apply the following strategy [3]. The training procedure is started with a simple model for which the problem of local optima does not occur or is not critical. In particular, the model IBM-1 has the advantage that it has only a single optimum and thus convergence problems cannot exist [3]. The parameters of the simple model are then used to initialize the training procedure of a more complex model. In such a way, a series of models with increasing complexity can be trained [10].

5. Generating the Target Sentence

The task of the search algorithm is to generate the most likely target sentence e_1^I of unknown length I for a source sentence f_1^J. The search must make use of all three knowledge sources as illustrated in Fig. 2: the alignment model, the (bilingual) lexicon model and the language model. All three of them must contribute in the final decision about the words in the target language. We replace the sum over all alignments by the best alignment, which is referred to as maximum approximation in speech recognition. Using a trigram language model $p(e_i|e_{i-2}^{i-1})$ and dropping the length model $p(J|I)$, we obtain the following search criterion:

$$\max_{I, B_1^I, e_1^I} \left\{ \prod_{i=1}^{I} \Big[p(e_i|e_{i-2}^{i-1}) \cdot p(B_i|B_1^{i-1}) \cdot \prod_{j \in B_i} p(f_j|e_i) \Big] \right\}$$

Considering this criterion, we can see that we can build up hypotheses of partial target sentences in a *sequential* strategy over the positions $i = 1, ..., I$ of the partial target sentence e_1^i. An important constraint for the alignment is that *all* positions of the source sentence should be covered exactly *once*. This constraint is similar to that of the travelling salesman problem where each city has to be visited exactly once [5].

The type of language model we use ranges from a trigram to a fivegram, which can be either word- or class-based. Beam search is used to handle the huge search space. To normalize the costs (i.e. the negative logarithm of the probabilities) of partial hypotheses covering different parts of the input sentence, an (optimistic) estimate of the remaining cost is added to the current accumulated cost as follows. For each word in the source sentence, a lower bound on its translation cost is determined beforehand. Using this lower bound, it is possible to achieve an efficient estimation of the remaining cost. Details on various search strategies can be found in [2, 12, 17, 19].

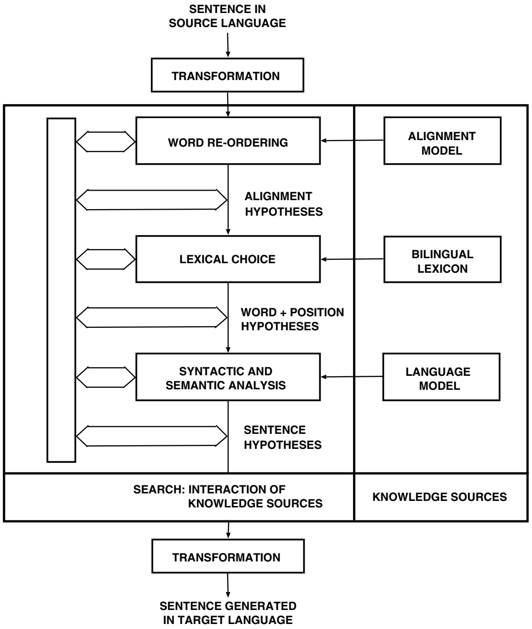

Figure 2: Generation of target sentence.

6. Progress and Improvements

6.1. Past Progress

In comparison with the baseline models developed in [2, 3], we have made the following extensions:

- The alignment model IBM-2 makes use of *absolute* word positions. We find that instead a HMM-type model works much better, in particular when it uses *relative* word positions [10].
- In order to obtain improved alignments in training, we symmetrize the training by exchanging the roles of source and target languages [10].
- For the generation of the target sentence, we go beyond the Model IBM-3 [2]. In addition, we allow more global word re–ordering [17], which of course is very much dependent on the language pair.
- For the search process to generate the target sentence, a dynamic programming beam search strategy is much more efficient than a (pure) A^* strategy [12, 17].
- The translation quality is significantly improved by modelling *word groups* rather than *single words* in both the alignment and the lexicon models [11]. The method is referred to as *alignment templates*.
- For some languages, e.g. German, it is useful to explicitly include some first steps towards a morpho-syntactic analysis [8].

6.2. Future Improvements

To achieve further improvements in statistical language translation, we consider the following methods to be promising:

- *grammar–based language models:* To improve the syntactic structure of the target sentences, we propose to study the use of grammar–based language models. Recently, there have been significant advances in the area of grammar models by probabilistic lexicalized context free grammars.
- *syntactic and morpho-syntactic information:* The syntactic structure, in combination with some morpho-syntactic analysis, should be taken into account for both target and source sentences. In such a way, we believe that the difference in the word order between target and source sentences can be better taken into account.
- *context dependent lexicon model:* The present approaches to statistical models for alignment and lexicon are fairly independent of 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.
- *improved algorithms of statistical and machine learning:* The past experience with speech and language processing has 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 work along these lines will result in significant improvements. In particular, promising directions are maximum entropy models and the use of discriminative criteria for training so that the learning can be directly aimed at optimizing the translation accuracy.
- *integration with speech recognition:* In the traditional approach to spoken language translation, only the first best sentence hypothesis produced by the speech recognizer is passed on as input to the translation component. An alternative method is to generate n-best lists of sentences or word lattices by the speech recognizer to allow for recognition errors. So the question comes up of how to integrate the probabilities of the speech recognition process into the translation process. From a strictly statistical point of view, we obtain a modified Bayes decision rule that integrates recognition and translation into a *single criterion* [6]. Although there has been already a lot of work on speech translation, this integrated approach has not been fully studied yet.

Acknowledgment. The work reported here was supported partly by the Verbmobil project (German BMBF) and as part of various European projects: Eutrans, LC-Star, PF-Star and TC-Star-P.

7. References

[1] 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.

[2] A. L. Berger, P. F. Brown, J. Cocke et al.: The Candide System for Machine Translation. *ARPA Human Language Technology Workshop*, Plainsboro, NJ, Morgan Kaufmann Publishers, San Mateo, CA, pp. 152-157, March 1994.

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

[4] F. Casacuberta, D. Llorenz, C. Martinez et al.: Speech-To-Speech Translation Based on Finite-State Transducers. *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, pp. 613-616, May 2001.

[5] K. Knight: Decoding Complexity in Word-Replacement Translation Models. *Computational Linguistics*, No. 4, Vol. 25, pp. 607–615, 1999.

[6] 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.

[7] H. Ney, F. J. Och, S. Vogel: The RWTH System for Statistical Translation of Spoken Dialogues. *Human Language Technology Conference*, San Diego, CA, pp. 302-308, March 2001.

[8] S. Nießen, H. Ney: Improving SMT Quality with Morpho-Syntactical Analysis. *Int. Conf. on Computational Linguistics 2000 (Coling)*, Saarbrücken, pp. 1081-1085, Aug. 2000.

[9] NIST evaluation conditions: `http://www.nist.gov/speech/tests/mt/mt2001/resource/`, June 2002.

[10] F. J. Och, H. Ney: A Systematic Comparison of Various Alignment Models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19-51, March 2003.

[11] 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*, pp. 20–28, University of Maryland, College Park, MD, June 1999.

[12] F. J. Och, N. Ueffing, H. Ney: An Efficient A* Search Algorithm for Statistical Machine Translation. *Data-Driven Machine Translation Workshop*, 39th Annual Meeting of the Ass. for Computational Linguistics (ACL), pp. 55–62, Toulouse, July 2001.

[13] S. D. Richardson, W. B. Dolan, A. Menezes et al.: Overcoming the customization bottleneck using example-based MT. *Data-Driven Machine Translation Workshop, 39th Annual Meeting of the Assoc. for Computational Linguistics*, pp. 9–16, Toulouse, France, July 2001.

[14] E. Sumita: Example-based Machine Translation using DP-Matching between Word Sequences. *Data-Driven Machine Translation Workshop, 39th Annual Meeting of the Assoc. for Computational Linguistics*, pp. 1–8, Toulouse, France, July 2001.

[15] E. Sumita, Y. Akiba, T. Doi et al.: A Corpus-Centered Approach to Spoken Language Translation. *10th Conf. of the Europ. Chapter of the Ass. for Computational Linguistics (EACL)*, Budapest, Hungary, pp. 171-174 (Conference Companion), April 2003.

[16] L. Tessiore, W. v. Hahn: ‘Functional Validation of a Machine Translation System: Verbmobil’, in [18], pp. 611-631.

[17] 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.

[18] W. Wahlster (Ed.): *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer-Verlag, Berlin, 2000.

[19] Y.-Y. Wang, A. Waibel: Decoding Algorithm in Statistical Translation. *35th Annual Conf. of the Association for Computational Linguistics*, pp. 366-372, Madrid, Spain, July 1997.

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

[21] 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.