Stochastic Modelling: From Pattern Classification to Language Translation

Hermann Ney
Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology
D-52056 Aachen, Germany
ney@informatik.rwth-aachen.de

Abstract

This paper gives an overview of the stochastic modelling approach to machine translation. Starting with the Bayes decision rule as in pattern classification and speech recognition, we show how the resulting system architecture can be structured into three parts: the language model probability, the string translation model probability and the search procedure that generates the word sequence in the target language. We discuss the properties of the system components and report results on the translation of spoken dialogues in the VERBMOBIL project. The experience obtained in the VERBMOBIL project, in particular a large-scale end-to-end evaluation, showed that the stochastic modelling approach resulted in significantly lower error rates than three competing translation approaches: the sentence error rate was 29% in comparison with 52% to 62% for the other translation approaches.

1 Introduction

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 correct 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 language 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:

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

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

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

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 modelled 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 which 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 modelling 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 modelling process and to enable improved generalization with respect to unseen data. Therefore in a good stochastic modelling 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.

2 Language Translation as Pattern Classification

2.1 Bayes Decision Rule

Knowing that language translation is a difficult task, we want to keep the number of wrong translations as small as possible. The corresponding formalism is provided by the so-called statistical decision theory. The resulting decison rule is referred to as Bayes decision rule and is the starting point for many techniques in pattern classification (Duda et al. 2001). To classify an observation vector y into one out of several classes c, the Bayes decision rule is:

$$\begin{aligned}\hat{c} &= \arg\max_{c} \{Pr(c|y)\} \\ &= \arg\max_{c} \{Pr(c) \cdot Pr(y|c)\} \ .\end{aligned}$$

For language translation, the starting point is the observed sequence of source symbols $y = f_1^J = f_1 ... f_J$, i.e. the sequence of source words, for which the target word sequence $c = e_1^I = e_1 ... e_I$ has to be determined. In order to minimize the number of decision errors at the sentence level, we have to choose the sequence of target words $\hat{e}_1^I$ according to the equation (Brown et al. 1993):

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

Here, the posterior probability $Pr(e_1^I|f_1^J)$ is decomposed into the language model probability $Pr(e_1^J)$ and the string translation probability $Pr(f_1^J|e_1^I)$. Due to this factorization, we have two separate probability distributions which can be modelled and trained independently of each other.

Fig.1 shows the architecture that results from the Bayes decision theory. Here we have already taken into account that, in order to implement the string translation model, we will decompose it into a so-called alignment model and a lexicon model. As also shown in this figure, we explicitly allow for optional transformations to make the translation task simpler for the algorithm.

In total, we have the following crucial constituents of the stochastic modelling approach to language translation:

- There are two separate probability distributions or *stochastic knowledge sources*:
  - the language model distribution $Pr(e_1^I)$, which is assigned to each possible target word sequence e_1^I and which ultimately captures all syntactic, semantic and pragmatic constraints of the target language domain under consideration;
  - the string translation probability distribution $Pr(f_1^J|e_1^I)$ which assigns a score as to how well the source string f_1^J matches the hypothesized target sequence e_1^I.
- In addition to these two knowledge sources, we need another system component which is referred to as a search or decision process. According to the Bayes decision rule, this search has to carry out the maximization of the product of the two probability distributions and thus ensures an optimal interaction of the two knowledge sources.

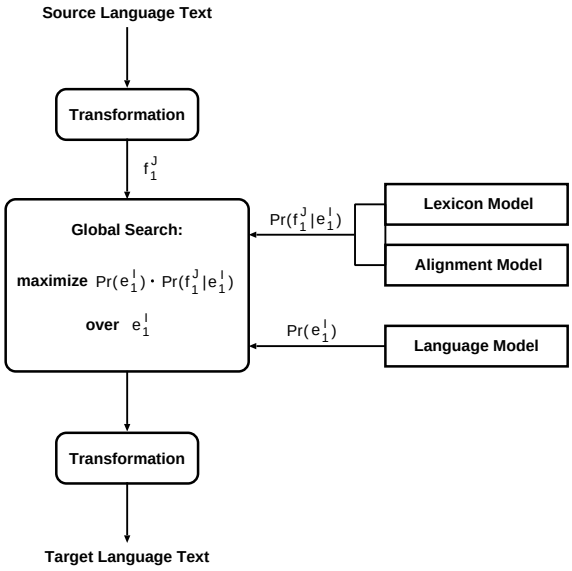

Figure 1: Bayes architecture for language translation.

Note that there is a *guarantee* of the minimization of decision errors if we know the *true* probability distributions $Pr(e_1^I)$ and $Pr(f_1^J|e_1^I)$ and if we carry out a *full* search over all target word sequences e_1^I. In addition, it should be noted that both the sequence of source words f_1^J and the sequence of unknown target words e_1^I are modelled as a whole. The advantage then is that context dependencies can be fully taken into account and the syntactic analysis of both source and target sequences (at least in principle) can be integrated into the translation process.

2.2 Implementation of Stochastic Modelling

To build a real operational system for language translation, we are faced with the following three problems:

- **Search problem:**
  In principle, the innocent looking maximization requires the evaluation of $20\,000^{10} = 10^{43}$ possible target word sequences, when we assume a vocabulary of 20 000 target words and a sentence length of $I = 10$ words. This is the price we have to pay for a full interaction between the language model $Pr(e_1^I)$ and and the string translation model $Pr(f_1^J|e_1^I)$. In such a way, however, it is guaranteed that there is no better way to take the decisions about the words in the target language (for the given probability distributions $Pr(e_1^I)$ and $Pr(f_1^J|e_1^I)$). In a practical system, we of course use suboptimal search strategies which require much less effort than a full search, but nevertheless should find the global optimum in virtually all cases.

- **Modelling problem:**
  The two probability distributions $Pr(e_1^I)$ and $Pr(f_1^J|e_1^I)$ are too general to be used in a table look-up approach, because there is a huge number of possible values f_1^J and e_1^I. Therefore we have to introduce suitable structures into the distributions such that the number of free parameters is drastically reduced by taking suitable data dependencies into account.

  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 $(f_1^J; e_1^I)$. Typically, the dependence is further constrained by assigning each source word to *exactly one* target word. Models describing these types of dependencies are referred to as *alignment mappings* (Brown et al. 1993):

  $$\text{alignment mapping:} \quad j \rightarrow i = a_j \ ,$$

  which assigns a source word f_j in position j to a target word e_i in position $i = a_j$. As a result, the string translation probability can be decomposed into a lexicon probability and an alignment probability (Brown et al. 1993).

- **Training problem:**
  After choosing suitable models for the two distributions $Pr(e_1^I)$ and $Pr(f_1^J|e_1^I)$, there remain free parameters that have to be learned from a set of training observations, which in the statistical terminology is referred to as *parameter estimation*. For several reasons, especially for the interdependence of the parameters, this learning task typically results in a complex mathematical optimization problem the

details of which depend on the chosen model and on the chosen training criterion (such as maximum likelihood, squared error criterion, discriminative criterion, minimum number of recognition errors, ...).

In conclusion, *stochastic modelling as such* does not solve the problems of automatic language translation, but defines a basis on which we can find the solutions to the problems. In contradiction to a widely held belief, a stochastic approach may very well require a specific model, and statistics helps us to make the best of a given model. Since undoubtedly we have to take decisions in the context of automatic language processing (and speech recognition), it can only be a rhetoric question of whether we should use statistical decision theory at all. To make a comparison with another field: in constructing a power plant, it would be foolish to ignore the principles of thermodynamics!

As to the search problem, the most successful strategies are based on either *stack decoding* or A^* search and *dynamic programming beam search*. For comparison, in speech recognition, over the last few years, there has been a lot of progress in structuring the search process to generate a compact *word lattice* or *word graph*.

To make this point crystal clear: The characteristic property of the stochastic modelling approach to language translation is *not* the use of *hidden Markov models* or *hidden alignments*. These methods are only the time-honoured methods and successful methods of today. The characteristic property lies in the systematic use of a probabilistic framework for the construction of models, in the statistical training of the free parameters of these models and in the explicit use of a global scoring criterion for the decision making process.

3 Experimental Results

Whereas stochastic modelling is widely used in speech recognition, there are so far only a few research groups that apply stochastic modelling to language translation (Berger et al. 1994; Brown et al. 1993; Knight 1999). 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).

We will consider the experimental results obtained in the VERBMOBIL project. 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 statistical approach, 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 et al. 2000). In each session of this evaluation, two native speakers conducted a dialogue. The speakers 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 neighbour 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 translation results using the following criterion: *Is the sentence approximatively 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

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

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 with the other approaches.

In agreement with other evaluation experiments, these experiments show that the statistical modelling approach may be comparable to or better than the conventional rule-based approach. In particular, the statistical approach seems to have the advantage if robustness is important, e.g. when the input string is not grammatically correct or when it is corrupted by recognition errors.

4 Conclusion

In summary, in the comparative evaluations, both text and speech input were translated with good quality on the average by the statistical approach. Nevertheless, there are examples where the syntactic structure of the produced target sentence is not correct. Some of these syntactic errors are related to long range dependencies and syntactic structures that are not captured by the m-gram language model used. To cope with these problems, morpho-syntactic analysis and grammar-based language models are currently being studied.

Acknowledgment

This paper is based on work supported partly by the VERBMOBIL project (contract number 01 IV 701 T4) by the German Federal Ministry of Education, Science, Research and Technology and as part of the EUTRANS project (ESPRIT project number 30268) by the European Community.

References

A. L. Berger, P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, J. R. Gillett, J. D. Lafferty, R. L. Mercer, H. Printz, L. Ures: "The Candide System for Machine Translation", *ARPA Human Language Technology Workshop*, Plainsboro, NJ, Morgan Kaufmann Pub., San Mateo, CA, pp. 152-157, March 1994.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer: "Mathematics of Statistical Machine Translation: Parameter Estimation", *Computational Linguistics*, Vol. 19.2, pp. 263-311, June 1993.

N. Chomsky: "Quine's Empirical Assumptions", in D. Davidson, J. Hintikka (eds.): *Words and objections. Essays on the work of W. V. Quine*, Reidel, Dordrecht, The Netherlands, 1969.

F. Casacuberta, D. Llorenz, C. Martinez, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Pico, A. Sanchis, E. Vidal, J. Vilar: "Speech-To-Speech Translation Based on Finite-State Transducers", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, May 2001.

R. O. Duda, P. E. Hart, D. G. Stork: *Pattern Classification*, 2nd ed., John Wiley & Sons, New York, NY, 2001.

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

H. Ney, F. J. Och, S. Vogel: "The RWTH System for Statistical Translation of Spoken Dialogues", *Human Language Technology Conference*, San Diego, CA, Proceedings in press, March 2001.

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.

L. Tessiore, W. v. Hahn: "Functional Validation of a Machine Interpretation System: Verbmobil", pp. 611–631, in (Wahlster 2000).

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