A Statistical-Estimation Method for Stochastic Finite-State Transducers Based on Entropy Measures

David Picó and Francisco Casacuberta

Institut Tecnològic d'Informàtica,
Departament de Sistemes Informàtics i Computació,
Universitat Politècnica de València
Camí de Vera, s/n, 46071 Valencia, Spain
{dpico,fcn}@iti.upv.es

Abstract The stochastic extension of formal translations constitutes a suitable framework for dealing with many problems in Syntactic Pattern Recognition. Some estimation criteria have already been proposed and developed for the parameter estimation of Regular Syntax-Directed Translation Schemata. Here, a new criterium is proposed for dealing with situations when training data is sparse. This criterium is based on entropy measurements, somehow inspired in the Maximum Mutual Information criterium, and it takes into account the possibility of ambiguity in translations (i.e., the translation model may yield different output strings for a single input string.) The goal in the stochastic framework is to find the most probable translation of a given input string. Experiments were performed on a translation task which has a high degree of ambiguity.

Keywords Machine translation, stochastic finite-state transducers, probabilistic estimation.

1 Introduction

A *translation* is a process that maps strings from a given language (the *input* language) into strings which belong to another language (the *output* language). If both the input and the output languages are formal, then the formal devices that implement such translations (*formal* translations) are known as *formal transducers* and have been thoroughly studied in the theory of formal languages [3]. Formal translations of many kinds were initially proposed for compiling programming languages [1] and as a framework for a concise presentation of error-correction models in syntactic pattern recognition [14].

Regular Translations constitute an important class of formal translations that have recently become of great interest as a model in some practical Syntactical Pattern Recognition problems in which the classification paradigm is not adequate [19] since the number of classes could be large or even infinite. In this case, the most general paradigm of *interpretation* seems to be a better framework and

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 417–426, 2000.

can be tackled through formal translations. For example, many tasks in Automatic Speech Recognition can be viewed as simple translations from acoustic sequences to sub-lexical or lexical sequences (Acoustic-Phonetic Decoding), or from acoustic or lexical sequences to sequences of commands to a data-base management system or to a robot (Semantic Decoding). A more complex application in the same line is the translation between natural languages (e.g., English to Spanish) [18,2]. Formal transducers can be learned automatically from examples [16,5]. This opens a wide field of applications based on the induction of translation models from parallel corpora.

However, the application of formal transducers to Syntactic Pattern Recognition needs a stochastic extension due to the noisy and distorted patterns which make the process of interpretation ambiguous [12]. The statistical parameters of the extended models define a probability distribution over the possible translations that help decide what is the best translation of a given input sentence. A common way of setting these parameters is to learn them from examples of translations.

A Maximum Likelihood algorithm for learning the statistical parameters of Stochastic Regular Syntax-Directed Translation Schemata from examples has recently been proposed [6,8]. This algorithm estimates the parameter set by maximizing the likelihood of the training data over the model. The Maximum Conditional Entropy estimation criterion which is presented in this paper (see Section 2) is based on some ideas from Maximum Mutual Information (MMI) [4,10] and can be particularly useful when the training data is sparse.

On the other hand, a learning algorithm based on the MMI criterion had been proposed in a previous paper [8]. We have found this criterium to be inadequate for translation. A discussion about this matter is given in Section 2.

2 The Translation Schema

Let Σ be an input alphabet and Δ be an output alphabet. A *Formal Translation* T can be defined as a subset of $\Sigma^\star \times \Delta^\star$. Note that respectively naming Σ and Δ as the *input* and *output* alphabets is an arbitrary decision. We will use the terms input and output whenever it helps to make the presentation clearer.

A *Regular Syntax-Directed Translation Schema* (RT) is defined as a tuple $T = (N, \Sigma, \Delta, R, S)$ in which N is a finite set of non-terminal symbols, Σ and Δ are finite sets of input and output terminal symbols and $S \in N$ is the initial symbol of the schema. R is a set of rules $A \to aB, zB$ or $A \to a, z$, where $A, B \in N$, $a \in \Sigma$ and $z \in \Delta^\star$.

A natural extension of the RT is given by the Stochastic Regular Syntax-Directed Translation Schema (SRT). An SRT is a pair (T, P), where T is an RT and $P : R \to]0, 1]$ is a function that assigns probability [1] values to the rules of

[1] For the sake of simplicity, in the remainder of the paper we will denote $Pr(X = x)$ as $Pr(x)$ and $Pr(Y = y|X = x)$ as $Pr(y|x)$ where X and Y are stochastic variables and x and y are two possible values of X and Y, whenever the correct meaning can be deduced from context.

the schema in such a way that the sum of probabilities of all rules rewriting a non-terminal A is equal to 1 (*proper* SRT [14]). Formally, for any non-terminal A_i, the set $\{A_i \to \beta_1, A_i \to \beta_2, \ldots, A_i \to \beta_n\}$ of all rules that rewrite A_i must satisfy the following condition of stochasticity:

$$\sum_{j=1}^{n} P(A_i \to \beta_j) = 1. \tag{1}$$

A finite sequence of rules $tf = (r_1, r_2, \ldots, r_n)$ such that

$$(S, S) \overset{r_1}{\Rightarrow} (x_1 A_1, y_1 A_1) \overset{r_2}{\Rightarrow} (x_1 x_2 A_2, y_1 y_2 A_2) \cdots \overset{r_n}{\Rightarrow} (x, y)$$

is known as a *translation form* for the *translation pair* $(x, y) \in \Sigma^\star \times \Delta^\star$. We will denote x as $input(tf)$, and y as $output(tf)$.

Each translation form is given a probability in the model. This probability is defined as the product of probabilities of all rules that are used in the translation form. I.e., given an RT T and a set of probabilistic parameters $\Phi(T)$ given by some function P, and given a translation form $tf = (r_1, r_2, \ldots, r_n)$:

$$Pr(tf|\Phi(T)) = P(r_1)P(r_2)\cdots P(r_n). \tag{2}$$

Remark that a translation pair $(x, y) \in \Sigma^\star \times \Delta^\star$ may be derived in T through more than one translation form (there may exist tf, tf' in T so that $input(tf) = input(tf') = x$ and $output(tf) = output(tf') = y$ and $tf \neq tf'$.) Thus, the *probability of a translation* (x, y) must be defined as the sum of probabilities of all translation forms that produce (x, y):

$$Pr(x, y|\Phi(T)) = \sum_{\substack{\forall tf / input(tf)=x \\ \wedge\, output(tf)=y}} Pr(tf|\Phi(T)). \tag{3}$$

Given a sentence x in the input language, how can we obtain a translation of x in the output language? Note that since the SRT can be ambiguous, a single input sentence may be mapped by the model into more than one output sentence. This is one of the reason why we need a stochastic extension of the models: we will use the statistical information in the model for deciding which of the many possible translations of an input sentence is the best one. Following this idea, we define the *stochastic translation* of an input string $x \in \Sigma^\star$ in a SRT T as the string $y^\star \in \Delta^\star$ into which x is translated with the highest probability:

$$y^\star = \underset{y \in \Delta^\star}{\operatorname{argmax}}\, Pr(y|x, \Phi(T)), \tag{4}$$

where

$$Pr(y|x, \Phi(T)) = \frac{Pr(x, y|\Phi(T))}{Pr(x|\Phi(T))}.$$

Given that the probability $Pr(x|\Phi(T))$ does not depend upon the maximization index y, we can rewrite (4) as:

$$y^{\star} = \underset{y \in \Delta^{\star}}{\operatorname{argmax}} Pr(x, y|\Phi(T)). \tag{5}$$

Computing the stochastic translations of the input sentences is the proper way to make a translation with an SRT. However, the calculation of (5) has been demonstrated to be an NP-hard problem [9]. The only possible algorithmic solution is the use of some variant of the $A^{\star}$ algorithm (e.g., the Stack-Decoding [15]), which presents exponential computational costs in the worst case and, therefore, may not be feasible for some real applications.

A computationally cheaper approximation to the stochastic translation can be defined. Instead of defining the probability of a translation as shown in (3), we will work with the so-called *Viterbi probability of a translation*. This is defined as the probability of the translation form that most probably yields (x, y):

$$\hat{Pr}(x, y|\Phi(T)) = \max_{\substack{\forall tf/input(tf)=x \\ \wedge\, output(tf)=y}} Pr(tf|\Phi(T)). \tag{6}$$

The *approximate stochastic translation* $y^{\star\star}$ of an input sentence x is computed as an approximation to the stochastic translation defined in (5). Here, the Viterbi probability is used instead of the standard probability:

$$y^{\star\star} = \underset{y \in \Delta^{\star}}{\operatorname{argmax}} \hat{Pr}(x, y|\Phi(T)). \tag{7}$$

There exists a polynomial algorithm for calculating (7). This algorithm searches for the maximum probability translation form tf for a given input string x so that $input(tf) = x$.

2.1 Estimation through Entropy Measurements

The stochastic translation schema introduced in the previous section is a statistical model which can describe probability distributions over the universe of all possible pairs of input-output strings, $\Sigma^{\star} \times \Delta^{\star}$. The shape of these distributions depends both on the structure of each particular schema and on the set of probability parameters associated with the set of rules. Therefore, the process of building one of these schemata as a model of a certain probability distribution may be performed in two separate phases. First, a non-stochastic schema is generated, and second, a set of probability values for the rules in the schema is chosen. The generation of the structure can be done either manually or automatically (there are techniques for inferring RTs from examples; see [16,5]). Once a set of rules is given, the set of parameters will often need to be estimated from a representative sample[2] of translation pairs. This way we can obtain a stochastic schema that approximates the empirical probability distribution.

[2] A *sample* is any finite collection of translations with repetitions allowed (a multiset drawn from $\Sigma^{\star} \times \Delta^{\star}$.)

The problem of estimating the parameters of a model from a finite set of data has been thoroughly studied in Statistics. A well-known, general-purpose method for this is the so-called *Expectation-Maximization* method [11]. Here we will use the Baum-Welch algorithm, a more specific version of the Expectation-Maximization method which is suitable for estimating the probabilities of rules in a SRT. Thus, our process of estimation of the parameters of an SRT will be as follows. First of all, we define a function that depends both on the statistical parameters that we want to estimate and on the pairs of sentences in the training data. This function is designed to be sensitive to the *relevant* information in the sample, in such a way that higher values of the function correspond to better approximations of the model to reality. We will use the Baum-Welch algorithm or some variation of it for finding the optimal value of this function.

Maximum Likelihood Estimation (MLE) was proposed in [6] for estimating the statistical parameters of an SRT. It is based on the following assumption: it is supposed that the sample has been generated by a model that describes perfectly the real probability distribution. Under this assumption, the maximization of the likelihood of the training sample tends to make the distribution converge to the real one for increasingly large samples. The likelihood function to be maximized is:

$$R_{MLE}(\Phi(T)) = \prod_{(x,y)\in TS} Pr(x, y|\Phi(T)), \tag{8}$$

where $\Phi(T)$ is the set of parameters of SRT T and TS is a training sample.

In real applications, however, the amount of available data is far from being "large" in the theoretical sense, and MLE shows up to be a very poor method for estimation. The method that we are presenting in the next section is designed to make a better use of sparse data than MLE, and it is based on concepts such as conditional entropy and information channels.

The *entropy* $H(X)$ is a measure of the number of bits that are needed to specify the outcome of a random event X [17]. Intuitively, entropy can be understood as a plausible measure of the level of uncertainty in the event. It is defined follows:

$$H(X) = -\sum_{x} Pr(x) \log Pr(x)$$

Similarly, the *conditional entropy* of the random event X given the random event Y is a measure of the uncertainty in X given the outcome of Y:

$$H(X|Y) = -\sum_{x,y} Pr(x, y) \log Pr(x|y)$$

A statistical translation system can be interpreted as a bidirectional channel, where two sources or random events produce sentences in each of the languages involved, respectively, following the real probability distribution of sentences in either language. The translation channel performs a probabilistic mapping: it

sets the probability $Pr(x, y)$ for each pair (x, y) of sentences where x belongs to one of the languages and y to the other.

If we assume that a real translation between two languages can be properly described as one of these statistical translation systems, then our goal is to obtain a statistical model m which is as close as possible to the theoretical real system. Let $H_m(X|Y)$ stand for the conditional entropy of X given Y with respect to the probability distributions in model m. It has been demonstrated in [4] that the inequality $H_m(X|Y) \geq H(X|Y)$ always holds. Furthermore, the smaller the value of $H_m(X|Y)$, the more the distribution in model m resembles the real distribution. $H_m(X|Y)$ and $H(X|Y)$ are equal when the two distributions are the same. Therefore, we want to choose some model m that minimizes $H_m(Y|X)$. On the other hand, note that channel models are symmetrical with respect to the direction of translation. Hence, we might also want the model to minimize $H_m(X|Y)$. These two simultaneous goals can be achieved by looking for a model that minimizes the sum of both values, $H_m(X|Y) + H_m(Y|X)$.

Maximum Conditional Entropy Estimation. The criterium just mentioned can be used to estimate the parameter set of an SRT. First, notice that:

$$H(X|Y) + H(Y|X) = -\sum_{x,y} Pr(x, y) \log \frac{Pr_m^2(x, y)}{Pr_m(x)Pr_m(y)}.$$

For a given RT T together with a parameter set $\Phi(T)$, $Pr_m(x, y)$ is equal to $Pr(x, y|\Phi(T))$. Since we do not know the real probability distribution, $Pr(x, y)$, we must instead assume that the pairs (x, y) in our sample TS are representative and choose $\Phi(T)$ to minimize

$$-\sum_{(x,y)\in TS} \log \frac{Pr^2(x, y|\Phi(T))}{Pr(x|\Phi(T))Pr(y|\Phi(T))}. \tag{9}$$

Hence, *Maximum Conditional Entropy Estimation* (MCEE) is the estimation algorithm consisting in maximizing the following function:

$$R_{MCEE}(\Phi(T)) = \prod_{(x,y)\in TS} \frac{Pr^2(x, y|\Phi(T))}{Pr(x|\Phi(T))Pr(y|\Phi(T))} \tag{10}$$

The reestimation formulae for this function will be obtained through the application of an extension of the Baum theorem to rational functions due to Gopalakrishnan *et al.* [13]. Let Q_{TS}^{MCEE} be a transformation from the space $\Phi(T)$ into itself. Then, $\forall(A \to aB, zB) \in R$ we have:

$$Q_{TS}^{MCEE}(P(A \to aB, zB)) = \frac{P(A \to aB, zB)\left(\frac{\partial \log P_{MCEE}(\Phi(T))}{\partial P(A\to aB, zB)} + C\right)}{\sum_{a',z',B'} P(A \to a'B', z'B')\left(\frac{\partial \log P_{MCEE}(\Phi(T))}{\partial P(A\to a'B', z'B')} + C\right)} \tag{11}$$

where the numerator can be expanded using

$$P(A \to aB, zB)\frac{\partial \log P_{MCEE}(\Phi(T))}{\partial P(A \to aB, zB)}$$
$$= \sum_{(x,y)\in TS} \left(\frac{2}{Pr(x,y|\Phi(T))} P(A \to aB, zB) \frac{\partial Pr(x,y|\Phi(T))}{\partial P(A \to aB, zB)} \right.$$
$$- \frac{1}{Pr(x|\Phi(T))} P(A \to aB, zB) \frac{\partial Pr(x|\Phi(T))}{\partial P(A \to aB, zB)}$$
$$\left. - \frac{1}{Pr(y|\Phi(T))} P(A \to aB, zB) \frac{\partial Pr(y|\Phi(T))}{\partial P(A \to aB, zB)} \right) \quad (12)$$

and C is an admissible constant [13].

The first term in the sum is proportional to the expected number of times that the rule is used in the training set in the Maximum Likelihood reestimation approach and can be computed as in [6]. The second term is proportional to the expected number of times that the rule of the *input grammar* of T is used for parsing the set of inputs of the training translations. This input grammar is $G_i = (N, \Sigma, R_i, S, P_i)$, where, if $(A \to aB, zB) \in R$, then $(A \to aB) \in R_i$ and $P_i(A \to aB) = P(A \to aB, zB)$. The formulae obtained for a MLE with Stochastic Grammars can be used to compute this term [7]. Similarly, the third term is proportional to the expected number of times that the rule of the *output grammar* of T is used for parsing the set of outputs of the training translations. This grammar is $G_o = (N, \Sigma, R_o, S, P_o)$, where, if $(A \to aB, zB) \in R$, then $(A \to zB) \in R_o$ and $P_o(A \to aB) = P(A \to aB, zB)$. As for the second term, a simple modification of formulae in [7] can be used to compute the third term.

Discussion about Maximum Likelihood Estimation. A different method based on entropy measures was proposed in [8]. It was a straight-forward application of the *Maximum Mutual Information Estimation* (MMIE) by Brown [4] to stochastic translation schemata. In Brown's MMIE it is claimed that minimizing the conditional entropy $H(Y|X)$ is equivalent to maximizing the *mutual information*, $I(X;Y)$, since

$$H(Y|X) = H(X) - I(X;Y).$$

$H(X)$ represents the entropy of the source X and is supposed to be determined by some known language model and, therefore, fixed. However, this approximation is not adequate if (as it is our case) the language model that is being used is not independent from the translation model. When dealing with SRTs the probabilities of sources X and Y given by the model, $Pr_m(x) = Pr(x|\Phi(T))$ and $Pr_m(y) = Pr(y|\Phi(T))$, are a function of the set of parameters of the SRT, $\Phi(T)$. Therefore, they are *not* fixed during the estimation process and MMIE cannot be applied as proposed in [8].

MMIE could be used if an independent model for the probabilities of the input and the output sentences were given. Such a model could be, for instance,

a probabilistic model that represented information about the context of appearance of sentences within a line of discourse.

3 Experiments

Some experiments were carried out to compare MCEE with MLE. The selected task was the translation of Spanish sentences into English, as defined in project EuTrans-I [2]. The semantic domain of the sentences is restricted to tourist information, consisting in sentences that a hotel guest would address to a hotel receptionist at the information desk. A parallel corpus of paired Spanish-English sentences was artificially generated.

The structure of an SRT was inferred from the corpus by means of a new method for building finite-state transducers using regular grammars and morphisms [5]. The inferred SRT contained 490 non-terminal symbols and 1438 rules.

Training was done with 5 different series of training sets. Each series was composed of 10 mutually including sets of increasing size, containing 25, 50, 75, 100, 125, 150, 175, 200, 250 and 300 pairs, respectively. A set containing 500 different translation pairs was used for testing. The test set is disjoint to all training sets. All results were averaged over the 5 series of experiments.

The test set perplexity for these experiments is shown on the left of figure 3 and word error rate (WER) is shown on the right. Both measures turned up to be significantly better for MCEE for the smaller training sets, while ML get better results when training sets are greater.

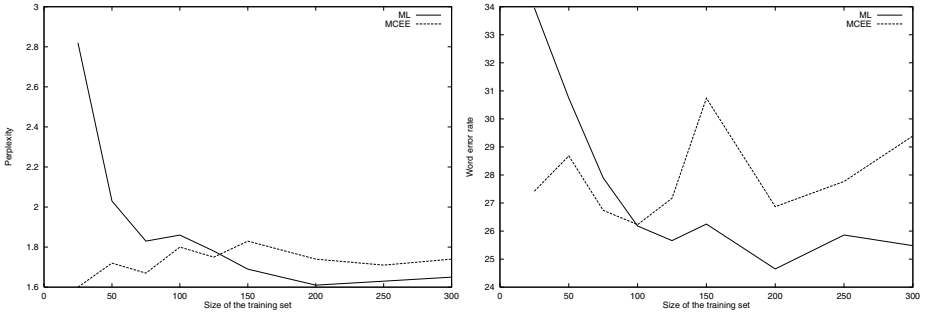

Figure 1. Test set perplexity and word error rate vs. size of the training set, respectively.

4 Conclusions

A new method for estimating the probabilistic parameters of an SRT with sparse training data has been presented in this paper. The method is based on the MMI

criterium, although it is different one, since direct application of MMI to SRTs is not adequate. Experiments on real data have been reported. MCEE exhibited better performance both in perplexity and word error rates for small training samples, while ML was better when the available amount of data was greater. This seems to point that MCEE is a good estimation criterium for the stochastic parameters of SRTs and may be specially useful when training data is scarce.

References

1. Aho, A. V. and Ullman, J. D. (1972). *The Theory of Parsing, Translation and Compiling*. Vol. 1. Prentice-Hall.
2. Amengual, J. C., Benedí, J. B., Casacuberta, F., Castaño, A., Castellanos, A., Jiménez, V. M., Llorens, D., Marzal, A., Pastor, M., Prat, F., Vidal, E., and Vilar, J. M. (1998). *The* EuTrans-I *Speech Translation System*. Submitted to Machine Translation.
3. Berstel, J. (1979). *Transductions and Context-Free Languages*. B. G. Teubner Stuttgart.
4. Brown, P. F. (1987). The Acoustic-Modelling Problem in Automatic Speech Recognition. Ph. Dissertation. Carnegie-Mellon University.
5. Casacuberta, F. (2000). Morphic Generator Translation Inference. To be submited for publication.
6. Casacuberta, F. (1995). Probabilistic Estimation of Stochastic Regular Syntax-Directed Translation Schemes. *Proc. of the VI Spanish Symposium on Pattern Recognition and Image Analysis*, pp. 201–207.
7. Casacuberta, F. (1996). Growth Transformations for Probabilistic Functions of Stochastic Grammars. *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 10, n. 3, pp. 183–201, Word Scientific Publishing Company.
8. Casacuberta, F. (1996). Maximum Mutual Information and Conditional Maximum Likelihood Estimation of Stochastic Regular Syntax-Directed Translation Schemes. *Grammatical inference: Learning Syntax from Sentences*, L. Miclet and C. de la Higuera (eds.). Lecture Notes in Artificial Intelligence. Vol.1147, pp. 282–291. Springer Verlag.
9. Casacuberta, F., de la Higuera, C. (1998). Computational Complexity of Problems on Probabilistic Grammars and Transducers. To be published.
10. Cardin, R., Normandin, Y., DeMori, R. (1994). High Performance Connected Digit Recognition using Maximum Mutual Information Estimation. *IEEE Trans. on Speech and Audio Processing*, vol.2 (2), pp. 300–311.
11. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society*, ser. B, vol. 39, num. 1, pp. 1–38.
12. Fu, K. S. (1982). *Syntactic Pattern Recognition and Applications*. Ed. Prentice-Hall.
13. Gopalakrishnan, P. S., Kanevsky, D., Nádas, A., and Nahamoo. D. (1991). An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems. *IEEE Transactions on Information Theory*, vol. 37, no. 1.
14. González, R. C., and Thomason, M. G. (1978). *Syntactic Pattern Recognition: An Introduction*, Addison-Wesley.
15. Jelinek, F. (1998) Statistical Methods for Speech Recognition. MIT Press, 1998.

16. Oncina, J., García, P., and Vidal, E. (1993). Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.15, no.5, pp. 448–458.
17. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, pp. 379–423 (Part I), pp. 623-656 (Part II).
18. Vidal, E. (1997). Finite-State Speech-to-Speech Translation. *Proceedings of the International Conference on Acoustic, Speech and Signal Processing*, vol.1, pp. 111-114. Munich (Germany).
19. Vidal, E., Casacuberta, F., and García, P. (1995). Grammatical Inference and Speech Recognition, *New Advances and Trends in Speech Recognition and Coding*. NATO ASI Series. pp. 174-191. Springer-Verlag.