AN ITERATIVE, DP-BASED SEARCH ALGORITHM FOR STATISTICAL MACHINE TRANSLATION [*]

Ismael García-Varea[1], *Francisco Casacuberta*[1], *and Hermann Ney*[2]

[1]Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
46071 Valencia, SPAIN

[2]Lehrstuhl für Informatik VI
RWTH Aachen, University of Technology
D-52056 Aachen, GERMANY

ABSTRACT

The increasing interest in the statistical approach to Machine Translation is due to the development of effective algorithms for training the probabilistic models proposed so far. However, one of the open problems with Statistical Machine Translation is the design of efficient algorithms for translating a given input string. For some interesting models, only (good) approximate solutions can be found. Recently a Dynamic Programming-like algorithm has been introduced which computes approximate solutions for some models. These solutions can be improved by using an iterative algorithm that refines the succesive solutions and uses a smoothing technique for some probabilistic distribution of the models based on an interpolation of different distributions. The technique resulting from this combination has been tested on the "Tourist Task" corpus, which was generated in a semi-automated way. The best results achieved were a word-error rate of 9.3% and a sentence-error rate of 44.4%.

1. INTRODUCTION

The statistical approach is an adequate framework for introducing automatic learning techniques in Machine Translation [3, 14, 5, 15].

Under this framework, given an input string **s** from $S^\star$ (S is a finite input alphabet and $S^\star$ is the set of finite length strings over S), the *probabilistic translation* of **s** is an output string, $\hat{\mathrm{e}} \in E^\star$ (E is a finite output alphabet) such that

$$\hat{\mathrm{e}} = \arg\max_{\mathbf{e}\in E^\star} \mathrm{Pr}(\mathbf{e}|\mathbf{s}) \tag{1}$$

Using Bayes' theorem, and taking into account that $Pr(\mathbf{s})$ is not a function of $\mathbf{e}$,

$$\hat{\mathrm{e}} = \arg\max_{\mathbf{e}\in E^\star} \mathrm{Pr}(\mathbf{s}|\mathbf{e})\mathrm{Pr}(\mathbf{e}) \tag{2}$$

Equation (2) is known as the **Fundamental Equation of Machine Translation** [4]. In Statistical Translation, the input string **s**, which is to be translated, is interpreted as a distorted string of an original string **e** from $E^\star$ through a noisy channel. The problem consist of finding and estimatie ê of the original string given the distorted string **s**. In this framework, Pr(**e**) represents the probability that the original string is produced, and Pr(**e**|**s**) is the probability that the original string **e** is distorted in the observed string **s**. The problem consist of finding an estimate ê of the original string given the distorted string **s**. In practice, an estimate of Pr(**e**) is used as a *Language Model* and an estimate of Pr(**s**|**e**) is used as a *Translation Model*. Among the reasons reported by Brown et al. [4] for using (2) instead (1), it can be observed that in (1), good output Language Models can aid the process of searching which allows for focusing on the *well-formed* output strings.

Interesting Translation Models were proposed in [4] and in [14]. With the model proposed in [14], a Dynamic Programming algorithm can be designed to solve (2) [10, 11]. However, the corresponding algorithms for the models 1 to 5 in [4] are based on a certain type of the $A^\star$ algorithm [3, 15]. The computational cost of this type of algorithms depends on the heuristics introduced. To overcome this problem we proposed in [6] a linear time approach on the total size of training data, that was based on a single Dynamic Programming-like algorithm which computes approximate solutions when the known IBM-Model2 from [4] is used. This proposal can be improved through an iterative process in which the solution is refined in succesive iterations, by using a technique similar to the one used in [6]. One of the problems of inferring statistical distribution from finite data is the problem of "unseen" events. So far, different techniques have been proposed for dealing with this problem [2] in language and in acoustic modeling. We have chosen an interpolation of a probabilistic distribution with different degrees of precision.

2. A STATISTICAL MODEL FOR MACHINE TRANSLATION

The Translation Models introduced by [4] are based on the concept of alignment between the components of the *translation pairs* (**s**,**e**) $\in S^\star \times E^\star$.

Formally, an alignment is a mapping between the sets of positions in **s** and **e**: **a** $\subset \{1, ..., |\mathbf{s}|\} \times \{1, ..., |\mathbf{e}|\}$. However, in [4], the concept of alignment is restricted to being a function **a**: $\{1, ..., |\mathbf{s}|\} \rightarrow \{0, ..., |\mathbf{e}|\}$, where $a_j = 0$ means that the position *j* in **s** is not aligned with any position of **e**. All the possible alignments between **e** and **s** are denoted by $\mathcal{A}$(**e**,**s**) and the probability of translating a given **e** into **s** by an alignment is denoted

[*]This work has been partially sopported by European Union under grant IT-LTR-O-30268

by Pr(**s**,**a**|**e**), therefore

$$\text{Pr}(\mathbf{s}|\mathbf{e}) = \sum_{\mathbf{a}\in\mathcal{A}(\mathbf{s},\mathbf{e})} \text{Pr}(\mathbf{s},\mathbf{a}|\mathbf{e}) \tag{3}$$

The second model for Pr(**s**,**a**|**e**) proposed in [4] (Model 2) is

$$\text{Pr}_{M2}(\mathbf{s},\mathbf{a}|\mathbf{e}) = \epsilon \prod_{j=1}^{|\mathbf{s}|} t(s_j \mid e_{a_j})\alpha(a_j|j;|\mathbf{s}|,|\mathbf{e}|) \tag{4}$$

where ϵ is a positive constant, $t(s_j \mid e_i)$ is the *translation probability* of the input word s_j given the output word e_i (by e_i we represent the i-th symbol of string **e** and by e_i^j the substring of **e** from i to j), and $\alpha(a_j|j;|\mathbf{s}|,|\mathbf{e}|)$ is the *alignment probability*. This distribution gives us the alignment probability of the i-th word in the target sentence, given any position in the source sentence and the length of both sentences. If equation (4) is used in (3), we have,

$$\text{Pr}_{M2}(\mathbf{s}|\mathbf{e}) = \epsilon \prod_{j=1}^{|\mathbf{s}|} \sum_{i=0}^{|\mathbf{e}|} t(s_j \mid e_i)\alpha(i|j;|\mathbf{s}|,|\mathbf{e}|) \tag{5}$$

Given $\{(\mathbf{s}^{(1)},\mathbf{e}^{(1)}),(\mathbf{s}^{(2)},\mathbf{e}^{(2)}),\cdots,(\mathbf{s}^{(K)},\mathbf{e}^{(K)})\}$, a training sample, the estimation of the translation probabilities and the alignment probabilities for both models can be performed by using the transformations proposed in [4]. These transformations allow us to increase the product of (5) for all training pairs in maximum likelihood training. In the case of IBM Model 1, the training procedure is guaranteed to find the global optimum.

3. SMOOTHING THE TRANSLATION MODEL

In both machine translation and speech recognition, we are often faced with the problem of estimating a large number of parameters from a relatively small amount of training data.

This is a typical problem in language modeling, above all when they are modeled with n-grams. To solve this problem, a lot of well-known techniques was proposed [7, 8, 9]. One of these techniques has been used for smoothing the distribution of alignment probabilities ($\alpha(i|j;|\mathbf{s}|,|\mathbf{e}|)$) shown in equation (5). The training data for these distributions are quite sparse due to the high number of parameters that are neccesary to estimate.

In order to solve this problem, the proposed smoothing technique is based on two simplifications of the model shown in (5):

- The Reduced Distribution, that consists of the elimination of the length of the source sentence, denoted by $\alpha_r(i|j;|\mathbf{e}|)$.
- The Simplified Distribution, that also eliminates the position of the source sentence in the last distribution. This is given by $\alpha_s(i||\mathbf{e}|)$.

The number of parameters required by these distributions is lower than the one in (5), and the modelling accuracy is lower too.

With these premises, the smoothing of the translation model is based on an interpolation [2] of the original probability distribution and the Reduced and Simplified distributions.

This can be seen as follows:

$$\begin{aligned} &\alpha_d(i|j;|\mathbf{s}|,|\mathbf{e}|) = \\ &\quad \beta * \alpha(i|j;|\mathbf{s}|,|\mathbf{e}|) + \gamma * \alpha_r(i|j;|\mathbf{e}|) \\ &\quad +(1-\beta-\gamma) * \alpha_s(i||\mathbf{e}|) \end{aligned} \tag{6}$$

Therefore equation (5) can be rewritten as:

$$\text{Pr}_{M2}(\mathbf{s}|\mathbf{e}) = \epsilon \prod_{j=1}^{|\mathbf{s}|} \sum_{i=0}^{|\mathbf{e}|} t(s_j \mid e_i)\alpha_d(i|j;|\mathbf{s}|,|\mathbf{e}|) \tag{7}$$

4. A PROCEDURE FOR MACHINE TRANSLATION

The goal of the Statistical Approach to MT is given by (2). The problem is to design an efficient algorithm for searching $\hat{s}$ (or an approximation to $\hat{s}$).

From (2)

$$\begin{aligned} &\max_{\mathbf{e}\in E^\star} \text{Pr}(\mathbf{e})\text{Pr}(\mathbf{s}|\mathbf{e}) = \\ &\qquad \max_{I} \max_{\mathbf{e}_1^I \in E^I} \text{Pr}(\mathbf{e}_1^I)\text{Pr}(\mathbf{s}_1^{|\mathbf{s}|} \,|\mathbf{e}_1^I) \end{aligned} \tag{8}$$

In other words, the maximisation in (2) can be performed by searching the best output string $\mathbf{e}_1^I$ for each possible I, and then by searching the optimal I.

Let us suppose that the length of the output string is known. The Translation Model used is (4) and, the Language Model will be a stochastic regular grammar, given by $G_R = (N,E,R,q_0,p)$, where: N is the set of non-terminals, E is the output alphabet, R is the set of rules like $q \rightarrow eq'$ or $q \rightarrow e$ (we suppose that exist a state $F \in N$ in order to allowing the last rule could be rewriten by $q \rightarrow eF$), q_0 is the first symbol of the grammar and p is a probabilistic function like $p : R \rightarrow]0,1]$ such that $\forall q \in N$:

$$\sum_{e\in E; q'\in N} p(q \rightarrow eq') = 1$$

For the sake of simplicity we will use $p(q_{i-1}, e_i, q_i)$ instead of $p(q_{i-1} \rightarrow e_i q_i)$, where q_i is the reached state when the symbol e_i was produced, begining in the state where e_{i-1} was produced, and so on.

From (5) and (8):

$$\begin{aligned} &\max_{\mathbf{e}_1^I\in E^I; q_1^I \in N^I} \left(Pr_{G_R}(\mathbf{e}_1^I) Pr_{M2}(\mathbf{s}_1^{|\mathbf{s}|} \mid \mathbf{e}_1^I)\right) = \\ &\quad \max_{\mathbf{e}_1^I\in E^I; q_1^I \in N^I} \Bigg(p(q_0,e_1,q_1) \prod_{i=2}^{I} p(q_{i-1},e_i,q_i) \\ &\quad \prod_{j=1}^{|\mathbf{s}|} \sum_{i=0}^{I} t(s_j|e_i)\alpha(i|j;|\mathbf{s}|,I) \Bigg) \end{aligned} \tag{9}$$

From (9), the following equation can be achieved: for $1 < i < I$ and $\forall e \in E$,

$$(\hat{q}_{i-1}(q_i), \hat{e}_i(q_i)) = \arg\max_{(q_{i-1}, e_i) \in N \times E} \Big(p(q_{i-1}, e_i, q_i) \times T(q_{i-1}, i-1) \times \prod_{j=1}^{|\mathbf{s}|} \big(Q(e_{i-1}, i-1, j) + t(s_j|e)\alpha(i|j; |\mathbf{s}|, I) + R(j, i+1)\big)\Big) \qquad (10)$$

where

$$R(j, i+1) = \sum_{k=i+1}^{I} t(s_j|\bar{e}_k)\alpha(k|j; |\mathbf{s}|, I) \qquad (11)$$

Note that $\bar{\mathbf{e}}_k$ $(i+1 \leq k \leq I)$ correspond to output symbols that are not yet explored, therefore, $\bar{\mathbf{e}}$ is obtained in an iterative process, i.e., $\bar{\mathbf{e}}$ is the guessed output ($\hat{\mathbf{e}}$) (the optimal one of a previous iteration). In the first iteration $R(j, i) = 0$ for $1 \leq i \leq I$ and $1 \leq j \leq |\mathbf{s}|$.

In the first iteration, no guessed output is used and the Q in (10) correspond to the approximate contribution of s_j to $\Pr_{M2}$ if the language model state q_i is achieved and e_1^i is produced. Thus,

$$Q(q_i, i, j) = \sum_{k=0}^{i} t(s_j|e_k)\alpha(k|j; |\mathbf{s}|, I) \qquad i > 1 \textit{ and } \forall j \qquad (12)$$

With

$$Q(q_1, 1, j) = t(s_j|e_0)\alpha(0|j; |\mathbf{s}|, I) + t(s_j|e)\alpha(1|j; |\mathbf{s}|, I) \qquad \forall j \qquad (13)$$

And T in(10) is

$$T(q_i, i) = \prod_{k=1}^{i} p(q_{k-1}, e_k, q_k) \qquad i > 1 \qquad (14)$$

With

$$T(q_1, 1) = p(q_0, e_1, q_1) \qquad (15)$$

From (10) to (15), $\forall e \in E$

$$T(q_i, i) = p(\hat{q}_{i-1}(q_i), \hat{e}_i(q_i), q_i) \times T(\hat{q}_{i-1}(q_i), i-1) \qquad (16)$$

and

$$Q(q_i, i, j) = Q(\hat{q}_{i-1}(q_i), i-1, j) + t(s_j|\hat{e}_i(q_i))\alpha(i|j; |\mathbf{s}|, I) \qquad (17)$$

The initialisations are (13) and (15).

Finally, the last symbol is obtained for $i = I$:

$$\hat{q}_I = \arg\max_{\forall q_I \in N} \left(T(q_I, I) \times \prod_{j=1}^{|\mathbf{s}|} Q(q_I, I, j) \right) \qquad (18)$$

The length of the output sentence is set statistically around the mean of the output lengths for each length of the input sentence.

The equations (11) to (18) are used in an iterative way:

```
algorithm TRANSLATION-SEARCH
  INPUTS
    - The translation probabilities $t$.
    - The alignment probabilities $\alpha$.
    - A general regular language model $G_R$.
    - An input string $\mathbf{s} \in S^\star$.
  OUTPUT
    - $\arg\max_e$ ($\Pr_{G_R}(\mathbf{e})$ $\Pr_{M2}(\mathbf{e}|\mathbf{s})$)
  METHOD
    - initialisation
        Compute a first approximation ($\hat{\mathbf{e}}$) to the solution by
        using equation (10) to (18) by setting $R(j, i) = 0$,
        $1 \leq i \leq I$ and $1 \leq j \leq |\mathbf{s}|$ and for the different output
        lengths $I$ according to the statistical distribution
        observed in the training set.
    - iteration
        While not convergence do
            Compute a new approximation ($\hat{\mathbf{e}}$) to the solution
            by using equations (10) to (18). In equation (11),
            $\bar{\mathbf{e}}$ corresponds to the $\hat{\mathbf{e}}$ of the previous iteration.
        end of While
end-of-algorithm
```

The computational time complexity of each iteration is $O(|\mathbf{s}| \times \mathbf{I_{max}} \times \mathbf{n_I} \times |\mathbf{E}|)$, where $\mathbf{I_{max}}$ is the maximum output length allowed and $\mathbf{n_I}$ is the number of output lengths tested.

5. EXPERIMENTS AND RESULTS

We selected the "Traveller Task" [13] to experiment with the search algorithm proposed here. The general domain of the task was a visit by a tourist to a foreign country. This domain included a great variety of different scenarios, from limited-domain applications to unrestricted natural language. The task used for the experiments reported here corresponded to a scenario of human-to-human communication situations at reception desk of a hotel. This task provided a small "seed corpus" from which a large set of sentence pairs was generated in a semi-automatic way [1]. From the different pairs of languages that were generated, only Spanish to English was considered for this work. The parallel corpus consisted of 500,000 sentence pairs (171,481 different sentence pairs). The input and output vocabulary sizes were 689 and 514, and the average input and output sentence lengths were 9.7 and 9.9, both respectively.

From the above corpus, a sub-corpus of 10,000 random sentence pairs was selected for training purposes. Testing was carried out with 500 input random sentences generated independently from the training set.

Under these circumstances two different experiments were done: Both used the task described above, the first one in its original form, and the other one categorizing certain words of the vocabulary, i.e. the proper names, dates, hours and numbers. In both cases, the algorithm was proven using the smoothed translation model.

The output language model was a Stochastic Regular Grammar built by the ECGI algorithm [12]. The output test-set perplexity of the inferred ECGI grammar was 3.53.

We tested the number of iterations for the proposed algorithm. There was no improvement in the word error rate when the number of iterations was increased beyond three. The results are shown in Tables 1 and 2 with and without smoothing of the alignment probabilities, respectively.

Error-Rate Percentage				
	1st Iter.		**3rd Iter.**	
Categories	**WER**	**SER**	**WER**	**SER**
NO	39.6 %	74.7 %	12.3 %	53.5 %
YES	37.7 %	65.3 %	10.1 %	46.3 %

Table 1: Translation results in first and third iterations of the algorithm. Word-Error Rate and Sentence-Error Rate for 500 test sentences. No smoothing was used for the alignment probabilistic distribution.

Error-Rate Percentage				
	1st Iter.		**3rd Iter.**	
Categories	**WER**	**SER**	**WER**	**SER**
NO	37.3 %	72.8 %	10.7 %	51.9 %
YES	36.3 %	64.4 %	9.3 %	44.4 %

Table 2: Translation results in first and third iterations of the algorithm. Word-Error Rate and Sentence-Error Rate for 500 test sentences. Smoothing was used for alignment probabilistic distribution, with a value of $\beta = 0.6$ and $\gamma = 0.3$.

6. CONCLUSIONS

A new iterative search algorithm for Statistical Translation has been proposed here. In the experiments, the Translation Model IBM-Model2 was combined with a Stochastic Regular Grammar under the search algorithm instead of the conventional bigram models. In this approach, all the components were learned automatically from training pairs; the Translation Model by a Maximum Likelihood Estimation procedure, and the Language Model by a Grammatical Inference technique. These techniques were tested on the "Traveller Task". The main conclusions that can be drawn are:

- Taking into account the complexity of the task, good results (as a Word-Error Rate and a Sentence-Error Rate measure of the translated sentences) can be achieved (with a linear time complexity algorithm).
- Comparing the Tables 1 and 2, slightly better results were obtained using the smoothing technique. In any case, the error rate decreased when lexical categories were used.
- As can be seen from Tables 1 and 2, the iterations drastically improve the translation quality.

7. REFERENCES

1. J. C. Amengual, J. M. Benedí, A. Castaño, A. Marzal, F. Prat, E. Vidal, J. M. Vilar, C. Delogu, A. di Carlo, H. Ney and S. Vogel. "Definition of a Machine Translation Task and Generation of Corpora". *Final Report, Part I. ESPRIT project No. 20268 EUTRANS*, 1996.
2. Lalit L. Bahl, Peter F. Brown, Peter V. de Souza, Robert L. Mercer and David Nahamoo. "A Fast Algorithm for Deleted Interpolation". In *Computational Linguistics*, 1209–1212. 1990.
3. P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, J. Jelinek, J. Lafferty, R. Mercer and P. Roosssina. "A Statistical Approach to Machine Translation". In *Computational Linguistics*, 16:79–85. 1990.
4. P. Brown, S. Della Pietras, V. Della Pietra and R. Mercer. "The Mathematics of Statistical Machine Translation: Parameter Estimation". In *Computational Linguistics*, 19:263–310. 1993.
5. I. Dan Melamed. "A Word-to-Word Model of Translational Equivalence". In *Procs. of the ACL97*. pp 490–497. Madrid Spain, 1997.
6. I. García-Varea and F. Casacuberta. "A Search Procedure for Statistical Translation". In *Procs. of the VII National Symposium on Pattern Recognition and Image Analysis*. pp 199–204. Barcelona Spain, 1997.
7. F. Jelinek, R.L. Mercer. "Interpolated Estimation of Markov Source Parameters from Sparse Data". In *Procs. of the Wokshop on Pattern Recognition in Practice*. pp 381–402. Amsterdam Holland, 1980.
8. S.M. Katz. "Estimation of Probabilities from Sparse Data for The Lenguage Model Component of a Speech Recognizer". In *IEEE Trans. on Acoustics, Speech and Signal Processiing*, Vol.ASSP-35, No.3, pp.400-401. 1987.
9. A. Nádas. "Estimation of Probabilities in the Language Model of the IBM Speech Recognition System". In *IEEE Trans. on Acoustics, Speech and Signal Processiing*, Vol.ASSP-32, No.4, pp.859-861. 1994.
10. C. Tillmann, S. Vogel, H. Ney and A. Zubiaga. "A DP based Search Using Monotone Alignements in Statistical Translation". In *Procs. of the ACL97*. pp 289–296. Madrid Spain, 1997.
11. C. Tillmann, S. Vogel, H. Ney and A. Zubiaga. "Accelerated DP based Search for Statistical Translation". In *Procs. of the EuroSpeech97*. Vol. 5, pp 2667–2670. Madrid Spain, 1997.
12. E. Vidal and N. Prieto. "Learning Language Models through the ECGI method". In *Speech Communication*, 11:299–309. 1992.
13. E. Vidal. "Finite-State Speech-to-Speech Translation". In *Procs. of the ICASSP97*, Vol.I, pp. 111–114. 1997.
14. S. Vogel, H. Ney and C. Tillmann. "HMM-Based Word Alignment in Statistical Translation". In *Procs. of the Int. Conf. on Computational Linguistics 1996*, pp. 836–841, Copenhagen Denmark Aug. 1996.
15. Y. Wang and A. Waibel "Decoding Algorithm in Statistical Machine Translation". In *Procs. of the ACL97*. pp 366–372. Madrid Spain, 1997.