Error Correcting Parsing for Text-to-text Machine Translation using Finite State Models

Juan C. Amengual[1,3] José M. Benedí[2] Francisco Casacuberta[2]
Asunción Castaño[1] Antonio Castellanos[1] David Llorens[2]
Andrés Marzal[1] Federico Prat[1] Enrique Vidal[2]
Juan M. Vilar[1]

(1) Unidad Predepartamental de Informática
Universidad Jaume I de Castellón
Campus de Penyeta Roja. Castellón, 12071. Spain
(2) Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera s/n. Valencia, 46071. Spain
(3) e-mail: jcamen@inf.uji.es

Abstract. This paper describes the approach followed to perform text-to-text Machine Translation (MT) in the first phase of the European project EUTRANS. This project aims at performing text and speech input MT in limited domain tasks. The EUTRANS system relies on Subsequential Transducers (SSTs), which are finite state translation models that can be automatically learned from training samples. Error Correcting Parsing is employed to increase the robustness of SSTs. After reviewing our approach, this paper presents results with the corpora defined within the EUTRANS project.

1 Introduction

The EUTRANS project, funded by the European Union, aims at developing Machine Translation (MT) systems for limited domain applications that require text and/or speech input, using *Example Based* techniques. This paper describes the techniques employed in the text-to-text translation system developed in this project. The basic translation models are Subsequential Transducers (SSTs), a kind of finite state models.

Despite the conceptual simplicity of SSTs, the results show that they can perform surprisingly well in limited domain tasks, that is tasks with small or medium sized vocabulary and restricted syntax [Vilar et al., 1997]. They also have the advantage of being learnable from training data using the Onward Subsequential Transducer Inference Algorithm (OSTIA) [Oncina et al., 1993]. An extension of OSTIA, known as OSTIA-DR [Oncina & Varó, 1996], allows to enforce syntactic constraints to the input and output languages of the learned SST.

The finite state nature of SSTs makes them amenable to being integrated with error models comprising insertions, substitutions and deletions [Vilar et al., 1997]. Error Correcting Parsing (ECP) is used not only to deal with input sentences that contain errors, but also to improve the performance achieved by SSTs with correct sentences.

The paper is organised as follows. In Section 2, the concept of SST is introduced and the basics of OSTIA and OSTIA-DR are outlined. Section 3 considers the use of ECP in order to improve the robustness of SSTs. In Section 4, the estimation procedures of the probabilities of language and error models are presented. Section 5 describes the experimental translation task chosen to test the performance of the EuTRANS MT system. Section 6 details the experiments performed and reports the results achieved by the system. Finally, some conclusions and future directions are presented in Section 7.

2 Subsequential Transducer Learning

A Subsequential Transducer [Berstel, 1979] is a deterministic finite state network that accepts sentences from a given input language and produces associated sentences of an output language. It is composed of states and edges connecting them. Each edge has associated an input symbol and an output string. The condition of determinism implies that no two distinct edges departing from a given state have the same input symbol. The processing of an input sentence begins from a distinguished state (the initial state) and proceeds by consuming input symbols one by one. Every time an input symbol is accepted, the string associated to the corresponding edge is output and a new state is reached. This process continues on until the whole input is processed; then, additional output may be produced from the last state reached in the analysis of the input.

A distinctive advantage of SSTs is the fact that they can be efficiently learned from unambiguous[1] training sets of input-output examples. This can be done by means of OSTIA [Oncina et al., 1993]. This algorithm basically works in three steps:

1. A finite state prefix tree acceptor is built from the input sentences. Then, empty strings are assigned as output substrings to the edges of this tree, while every output sentence is associated as a whole to the state reached by the corresponding input string. This is the initial SST.
2. The longest common prefixes of the output strings are recursively moved, level by level, from the leaf states of the tree towards the root.
3. Starting from the root state, all pairs of states are orderly considered, level by level, and they are merged if merging is *acceptable;* i.e., if the resulting transducer is Subsequential and is not in contradiction with the training set.

Usually, the SSTs learned by OSTIA constitute good translation models. However, since OSTIA does not take the syntactic structure of the input (domain) and output (range) languages into account, it often produces poor input language models. This provides highly accurate translations for correct input sentences, along with acceptance and meaningless translations for even slightly incorrect sentences. In practice, this can lead to very negative effects in case of imperfect input, as is expected with speech or spontaneous text input.

The algorithm OSTIA-DR [Oncina & Varó, 1996] incorporates domain and range models in the learning process. These models are deterministic finite state automata. OSTIA-DR allows to learn SSTs that accept only sentences compatible with the input model and produce only sentences compatible with the output model. These SSTs are better language models than those learned using OSTIA.

3 Translation using Error Correcting Parsing

SSTs cannot translate input sentences that do not comply with the syntactic restrictions imposed by them. This can arise from errors in the input sentences and lack of generalisation

[1] The training set contains no two pairs with the same input and different output.

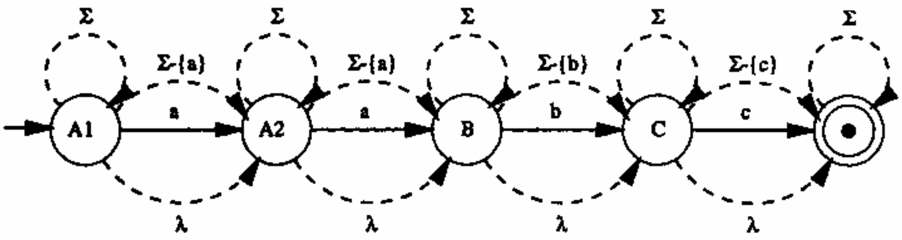

Figure 1. A finite state automaton, representing $L = \{aabc\}$, extended for including our error model, which comprises insertions (transitions labeled Σ), substitutions (transitions labeled $\Sigma-\{$correct transition symbol$\}$) and deletions (transitions labeled λ)

in learning. The input errors can have the form of vocabulary variations, repetitions, word disappearances, superfluous and misplaced words[2], and so on. Lack of generalisation leads to rejection of sentences that are correct with regard to the task under consideration.

In both cases, the use of ECP techniques can give the necessary robustness to SSTs in order to overcome these problems [Amengual et al., 1996b, Vilar et al., 1997]. Under this approach, the input sentence x_I is considered as a corrupted version of some sentence $x \in L$, being L the domain or input language of the SST. The corruption process is modeled by means of an *error model*, E, that accounts for insertions, substitutions and deletions. We assume that x minimises some dissimilarity function $d_E : \Sigma^* \times \Sigma^* \to \mathbb{R}$, where Σ stands for the input alphabet of the SST. An example of such a function is the Levenshtein edit distance [Kruskal, 1983].

In the case of stochastic ECP, we assume that the corruption process follows the probability distribution function $P_E(\cdot|\cdot)$, induced by E. Furthermore, assuming that a stochastic model for L is known, we will use $P_L(\cdot)$ to represent the probability distribution function induced by this model. Then, we estimate x by the sentence $\hat{x} \in L$ which maximises its *a posteriori* probability given x_I; equivalently,

$$\hat{x} = \underset{x' \in L}{argmax}\, P_L(x') \cdot P_E(x_I|x').$$

Finally, the translation returned for x_I is the translation of $\hat{x}$ by the SST.

When both E and the model of L are finite state stochastic models, they can be easily combined into an integrated finite state stochastic model using a simple construction [Fu, 1982]. An example of this construction can be seen in Figure 1 for our choice of E in the case of $L = \{aabc\}$. This extended automaton accepts any sentence $x \in \Sigma^*$, effectively providing a kind of smoothing. An analogous construction can be used in the case of L being a context free language [Fu, 1982].

Combined ECP and translation can be performed in order to find both a best error corrected sentence and its translation (given by the SST). The search problem can be solved very efficiently using fast Viterbi ECP techniques such as those proposed in [Amengual & Vidal, 1996].

4 Estimation of Language and Error Model Probabilities

When using stochastic ECP, once the structure of a model for L is learned by OSTIA or OSTIA-DR and a structure for E is fixed, their probabilistic parameters are estimated from training

[2] We interpret words as symbols, so vocabulary and alphabet are synonimous for us.

data. The parameters for the model of L are estimated by maximum likelihood from the input sentences of the training corpus used in the learning of the SST.

The parameters considered for E are:

—A probability of insertion for each symbol of Σ.
—A probability of deletion for each symbol of Σ.
—A probability of substitution for each symbol pair of $\Sigma \times \Sigma$.

These parameters are estimated using a corpus D which contains pairs consisting in a sentence x_C from the input language and a possibly erroneous version x_E The reestimation procedure of the parameters of E is done by repeating the following steps for each sentence:

- The sentence x_E is parsed using Viterbi ECP, obtaining a sentence x_L.
- The longest common subsequence between x_L and x_C is computed, let us call it l.

—The counts of use of the edges associated with l are stored.

Finally, the relative frequencies of use of error edges are used as estimates for the parameters of E. In each iteration, the parsing uses the parameters obtained in the previous iteration, except for the first iteration, when the Levenshtein edit distance [Kruskal, 1983] is used as dissimilarity function. This parsing and estimation process is repeated a fixed number of iterations, although other convergence criteria can be used [Amengual et al., 1996b, Vilar et al., 1997].

5 The Traveler Task

The task chosen to test the EUTRANS system was called the *Traveler Task,* a limited domain translation task in which a foreign traveler talks to a hotel receptionist. In the first phase of the project, the scenario has been limited to the following specific situations: notifying a previous reservation, asking about rooms (availability, features, price), having a look at rooms, complaining about and changing them, signing the registration form, asking for rooms, wake-up calls, keys, the bill, a taxi and moving the luggage, notifying the departure, asking and complaining about the bill, and other common expressions. Three text bilingual corpora (input: Spanish, output: English, German and Italian) were automatically generated using Stochastic, Syntax-directed Translation Schemata as described in [Amengual et al., 1996a]. Table 1 shows some examples of the resulting corpora. Table 2 summarises the main features of these three corpora[3].

6 Experiments and Results

For a preliminary evaluation of the possibility of recovering from imperfect input using ECP techniques, and given the impossibility of acquiring natural (imperfect) corpora during the first phase of EUTRANS, a distorted training corpus, composed of 64,000 sentences, was automatically derived from a subset of the input (Spanish) sentences of the original training corpora using a distortion model involving insertion, substitution and deletion errors [Hunt, 1988]. Three different percentages of global distortion—evenly distributed among each type of error—were used:

[3] Note that there are small discrepancies between these figures and those published elsewhere. This is due to the fact that the other figures were computed over a preliminary version of the corpora.

Table 1. Some examples of sentence pairs from the Traveler Task

Spanish:	*Por favor, llámenos a un taxi.*
English:	*Could you call a taxi for us, please?*
Spanish:	*Por favor, ¿podrían repasar la cuenta de la habitación seis cero tres ?*
German:	*Könnten Sie die Rechnung des Zimmers sechs null drei überprüfen, bitte ?*
Spanish:	*Por favor, reservamos una habitación doble para esta noche.*
Italian:	*Per favore, abbiamo prenotato una stanza doppia per questa notte.*

Table 2. Main features of the Spanish-English, Spanish-German and Spanish-Italian corpora

Feature	Spanish-English	Spanish-German	Spanish-Italian
Sentence pairs	500,000	500,000	500,000
Different sentence pairs	171,352	166,221	169,644
Input vocabulary size	689	691	687
Output vocabulary size	514	566	583
Average input length	9.5	8.9	12.7
Average output length	9.8	8.2	11.8

1%, 5% and 10%. An independent test set consisting of 1,000 Spanish sentences was distorted using the same percentages of global distortion.

For each of these three pairs of languages, a SST was trained with 490,000 training pairs (of which 168,629 were distinct in English, 163,505 in German and 166,897 in Italian). The training of these SSTs included the use of categories as explained in [Amengual et al., 1997]. Then, the probabilities of each SST were estimated by maximum likelihood from the original "clean" training corpus employed to learn the transducer. Next, the parameters of the error model were estimated from the distorted data following the reestimation procedure described in Section 4 (20 iterations). Finally, the distorted Spanish test sentences were submitted to translation using both conventional parsing and ECP, with the trained models. The translations obtained in this way were then compared with the target translations of the original 1,000 input sentences.

The results are shown in Table 3 in terms of translation Word Error Rate (WER), measured as the percentage of words that need to be inserted, deleted or substituted in order to obtain the correct translation. The first row of this table, labelled “No ECP”, corresponds to conventional parsing, while ECP results are reported in the following rows for increasing sizes of the distorted training set, ranging from 1,000 to 64,000 sentences. The power of ECP to deal with imperfect input sentences is clear from this table. Conventional parsing leads in large errors, even for the smallest distortion. In contrast, the error correcting analysis not only avoids these dramatic failures, but also achieves a significant recovery from errors when the error models have been adequately trained. For instance, for the better trained models and the largest input distortions (5% and 10%), output distortions around 1/3 of that of the input are achieved.

ECP is useful not only to deal with imperfect input, but also to improve the performance of imperfect SSTs for correct input sentences. For this purpose, the error model estimated from the 64,000 automatically distorted training sentences was also used (together with the same SSTs used in the above reported experiments) to translate the Spanish sentences of three different independent *undistorted* test sets (Spanish-English with 2,730 pairs, Spanish-German

Table 3. Experiments with imperfect input: translation Word Error Rate (in percentage) for different levels of global distortion *(Δ)* of the input sentences, without ECP, and with ECP for increasing number of *distorted* training sentences

Distorted	Translation WER								
Training	Spanish-English			Spanish-German			Spanish-Italian		
Sentences	*Δ=1%*	Δ=5%	Δ=10%	*Δ=1%*	*Δ=5%*	Δ=10%	Δ=1%	Δ=5%	Δ=10%
No ECP	13.06	47.75	74.53	13.62	49.61	76.39	12.33	48.00	74.09
1,000	4.87	7.88	12.46	5.21	8.56	12.45	2.95	6.37	9.82
2,000	3.23	5.59	9.11	3.48	5.91	9.27	2.02	4.97	7.62
4,000	2.17	3.94	7.18	2.11	4.08	7.23	1.47	4.11	6.29
8,000	1.58	3.07	5.82	1.56	3.07	6.11	1.16	3.54	5.63
16,000	0.66	1.93	4.39	1.12	2.59	4.96	0.93	2.79	4.99
32,000	0.66	1.82	3.81	1.05	2.34	4.64	0.89	2.54	4.30
64,000	0.65	1.68	3.48	1.04	2.20	4.55	0.89	2.31	4.25

Table 4. Experiments with perfect input: translation Word Error Rate (in percentage) without and with ECP

Test set	No ECP	ECP
Spanish-English	0.74	0.18
Spanish-German	1.23	0.54
Spanish-Italian	2.54	0.51

with 2,718 and Spanish-Italian with 2,751). The results, measuring the translation WER of the output translations with respect to the target translations, are shown in Table 4, together with the results obtained using conventional parsing (column "No ECP"). From these results, it is clear that ECP leads to a significant error reduction.

In addition, an experiment was performed with spontaneous sentences. In this case, the parameters of the stochastic error model were estimated using a dictionary of *synonyms* to produce the distorted training data. Therefore, only corrupted sentences having likely vocabulary variations were produced, rather than completely random distorted data. This way, the effective input vocabulary of the system is increased and errors similar to those expected in spontaneous language are properly accounted for by the trained model [Amengual et al., 1996b, Vilar et al., 1997]. Even in this case, it is possible that the users employ words not in the extended vocabulary. To account for this, the error model is smoothed by assigning a small probability to the generation, by means of an insertion or a substitution, of a special symbol representing out-of-vocabulary words.

To collect test data, several casual users were asked to write down a sentence for each situation described in Section 5. A total of 166 spontaneous Spanish sentences were so collected. The translations of the system into English were manually classified as belonging to one of these three categories[4]: *correct,* the translation preserves the meaning of the input sentence according to the definition of the task; *approximate*, the translation approximates the meaning of the input sentence according to the definition of the task; and *wrong*. The results were: correct, 68 sentences (40.96%); approximate, 51 sentences (30.72%); wrong, 47 sentences (28.31%).

[4] There were no expected translations for these sentences, thus making WER measures unfeasible.

Table 5. Translations of spontaneous sentences for the Traveler Task. “I” means input Spanish sentence, “C” means cleaned Spanish sentence and “O” means the translation of the cleaned sentence into English

I:	*Me gustaría que me avisasen a las sie*te *de la mañana pan ir de excursión.*
C:	*Querría que me despierten a ¡as siete de la mañana.*
Correct O:	*I would like you to wake me up at seven in the morning.*
I:	*Ruego que me cambien de habitación.*
C:	*Nos gustaría cambiarnos de habitación.*
Approx. O:	*We would like to change rooms* .
I:	*Me gustaría que me preparara la cuenta.*
C:	*Me parece que hay un problema en la cuenta.*
Wrong O:	I *think that there is a problem in the bill* .

Table 5 shows examples of translations that fall into each category. In the first example, **Me** was considered by the ECP as an insertion and **gustaría** was considered as a substitution of **Querría**, thus finding **Me gustaría** equivalent to **Querría**. Likewise, **avisasen** was considered as a substitution of **despierten**, which is adequate in the context defined by the task. Finally, the words **para ir de excursión** were considered as insertions, but since they contain no information according to the definition of the task[5], the sentence was classified as correct. In this example, **avisasen** and **excursión** are out-of-vocabulary words. In the second example, there is a mistake in the translation (**We** instead of **I**). However, the main request expressed in the input sentence is perfectly translated (**to change rooms**). The last example shows a completely wrong translation, where a request to prepare the bill is translated as a complaint about it.

7 Concluding Remarks and Future Directions

Automatic translation of unrestricted text is far from being satisfactorily solved. However, many applications of interest can be limited to a small or medium sized vocabulary, and have a restricted semantic domain. The use of SSTs is adequate for these kind of tasks. Their finite state nature makes them amenable to be used along with ECP techniques in order to improve their robustness and accuracy. The building of all these models can be done at a low cost since all of them can be automatically learned from training data. Specific techniques that reduce the amount of needed data can be applied, like the reordering of the output sentences proposed in [Vilar et al., 1997] and the ECP techniques presented here.

The feasibility of the approach for a real-world situation clearly relies on the availability of the required training data. A bootstrapping approach can be followed to obtain this corpus; starting with initial models, a translation system is built and submitted to practical use. Those sentences rejected or incorrectly translated are collected, leading to a first corrupted training set. This set is then used to re-train the error model which is used in turn to improve the future performance of the system, which is again submitted to practical use, and so on.

Also, the performance of ECP training could be enhanced by letting the updating procedure to be “human-guided” using *N*-Best hypotheses. These hypotheses could be automatically proposed by the error correcting parser itself, and could facilitate the choice of the adequate one; i.e., that which preserves the meaning of the original sentence.

[5] The receptionist has to wake the traveler up at seven in the morning, no matter what for.

8 Acknowledgements

Work mainly funded by the European Union EUTRANS project, IT-LTR-OS-20268, and partially funded by the Spanish CICYT TRACOM project, TIC95-0984-C02.

The authors wish to thank the anonymous reviewer for the careful reading and valuable comments.

References

[Amengual & Vidal, 1996] Amengual, Juan C. and Enrique Vidal. 1996. Different Approaches for Efficient Error-Correcting Viterbi Parsing: An Experimental Comparison. Research report DSIC-11/32/96, Dpto. de Sistemas Informáticos y Computación, Univ. Politécnica de Valencia, Spain.

[Amengual et al., 1996a] Amengual, Juan-Carlos, José-Miguel Benedí, Asunción Castaño, Andrés Marzal, Federico Prat, Enrique Vidal, Juan Miguel Vilar, Cristina Delogu, Andrea Di Carlo, Hermann Ney and Stephan Vogel. 1996. Definition of a Machine Translation Task and Generation of Corpora. Deliverable One, EUTRANS (IT-LTR-OS-20268). Restricted.

[Amengual et al., 1996b] Amengual, Juan C., Enrique Vidal and José M. Benedí. 1996. Simplifying Language through Error-Correcting Decoding. In *Proceedings of the Fourth International Conference on Spoken Language Processing* (ICSLP-96), Philadelphia, USA, pp. 841-844.

[Amengual et al., 1997] Amengual, J. C., J. M. Benedí, F. Casacuberta, A. Castaño, A. Castellanos, D. Llorens, A. Marzal, F. Prat, E. Vidal, and J. M. Vilar. 1997. Using Categories in the EUTRANS System. In *Proceedings of the Spoken Language Workshop* (SLT-97), Madrid, Spain.

[Berstel, 1979] Berstel, Jean. 1979. *Transductions and Context-free Languages.* Stuttgart: Teubner.

[Fu, 1982] Fu, King S. 1982. *Syntactic Pattern Recognition and Applications.* Engelwood Cliffs: Prentice Hall.

[Hunt, 1988] Hunt, Melvin J. 1988. Evaluating the performance of connected-word speech recognition systems. In *Proceedings of the 1988 International Conference on Acoustics, Speech, and Signal Processing* (ICASSP-88), New York, USA pp. 457-460.

[Kruskal, 1983] Kruskal, Joseph B. 1983. An Overview of Sequence Comparison. In D. Sankoff and J. B. Kruskal (eds) *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison,* Reading: Addison-Wesley, pp. 1-44.

[Oncina et al., 1993] Oncina, José, Pedro Garcia and Enrique Vidal. 1993. Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15:448-458.

[Oncina & Varó, 1996] Oncina, José and Miguel A. Varó. 1996. Using Domain Information During the Learning of a Subsequential Transducer. In L. Miclet and C. De La Higuera (eds) *Grammatical Inference: Learning Syntax from Sentences,* Lecture Notes in Artificial Intelligence 1147. Berlin: Springer-Verlag, pp. 301-312.

[Vilar et al., 1997] Vilar, Juan M., Víctor M. Jiménez, Juan C. Amengual, Antonio Castellanos, David Llorens and Enrique Vidal. 1997. Text and Speech Translation by means of Subsequential Transducers. To appear in *Journal of Natural Language Engineering.*