A Spanish-Catalan Translator Using Statistical Methods

Jesús Tomás [1] and **Jaime LLoret** [2] and **Francisco Casacuberta** [1]

Abstract. The development of a Spanish-Catalan statistical machine translation system is described in this paper. The methodology used attempts to solve the problem using a purely inductive approach, without using linguistic knowledge. To obtain the translator we use the following steps: First, we obtain a bilingual corpus from the internet. Second, we fragment the corpus into units (sentences and tokens). Third, we align the sentences from the two different languages and we use the aligned corpus to train statistical models. We use a trigram model as the target language model and a phrase-based model as translation model. Finally, we use these models to translate. In other words, given a source sentence, we search for the most probable target sentence. We have compared our translator with the most commonly used Spanish-Catalan translators and we have obtained similar translation results to the other commercial systems.

1 STOCHASTIC TRANSLATION

Today, machine translation systems are of ever-increasing interest in making communication among human beings easier. Different areas contribute to the development of such systems: linguistics, artificial intelligence and pattern recognition. Within the context of pattern recognition, the use of statistical methods seems to be a very promising approach.

The goal of statistical machine translation is to translate a given source language sentence $\mathbf{f} = f_1^{|\mathbf{f}|} = f_1 ... f_{|\mathbf{f}|}$ to a target sentence $\mathbf{e} = e_1^{|\mathbf{e}|} = e_1 ... e_{|\mathbf{e}|}$. Where $|\mathbf{f}|$ is the number of words in the source sentence and $|\mathbf{e}|$ is the number of words in the target sentence. The methodology used with stochastic translation [2] is based on the definition of a function $Pr(\mathbf{e}|\mathbf{f})$ that returns the probability of translating a given source sentence, $\mathbf{f}$, into a target sentence, $\mathbf{e}$. Once this function is estimated, the problem can be reduced to computing a sentence $\mathbf{e}$ that maximizes the probability $Pr(\mathbf{e}|\mathbf{f})$ for a given $\mathbf{f}$. Using Bayes' theorem, we can write:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} Pr(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e}} Pr(\mathbf{e}) Pr(\mathbf{f}|\mathbf{e}) \qquad (1)$$

Equation 1 summarizes the following three matters to be solved:

- An output language model is needed to distinguish valid sentences from invalid sentences in the target language, $Pr(\mathbf{e})$.
- A translation model, $Pr(\mathbf{f}|\mathbf{e})$.
- The design of an algorithm to search for the sentence $Pr(\mathbf{e})$ that maximizes this product. The search must be fast and efficient, even at the risk of a suboptimal result.

[1] Instituto de Tecnología Informática, Universidad Politécnica de Valencia, Valencia, Spain email: jtomas@upv.es
[2] Escuela Politécnica Superior de Gandia, Universidad Politécnica de Valencia, Spain

2 ACQUISITION OF TRAINING CORPORA

To be able to successfully translate a text using inductive techniques, it is necessary to have a large computerized database of parallel sentences. The automatic corpus construction process consists of three main phases:

Acquisition from Internet: For this purpose, we have automated the obtainment of two bilingual corpora [5]. The "DOGV" corpus is obtained from the "Diario Oficial de la Generalidad Valenciana" (http://www.gva.es/servic/predocas.htm) an official publication of the Valencian local government. And "El Periódico" corpus is obtained from the electronic publication of the newspaper "El Periódico de Cataluña" (http://www.elperiodico.es). This general information newspaper is published daily in a bilingual edition.

Text Fragmentation and Identifiable Translation Units: The aim of fragmentation is to break the source text up into linguistic fragments that can be viewed as units for translation purposes. In our project, we are interested in detecting articles, paragraphs, sentences and tokens. Some kinds of tokens cannot be enumerated in a dictionary, but can be detected by their structure. This is the case of tokens like 123,231 or MCIV that do not appear in the dictionary, but that are recognized as numbers. These kinds of tokens are known as identifiable translation units. In our system we are interested in detecting acronyms, abbreviations, numbers, years and proper nouns.

Alignment of Sentences in Parallel Texts: To identify the correspondence between sentences in one language and sentences in the other language. See [5] for more details about this phase.

We used these techniques to obtain two Spanish-Catalan corpora. The DOGV corpus was made up of 398 days of the publication (numbers 3,595 to 3,992). The "El Periódico" corpus was made up of 297 newspapers (from 1/7/99 to 31/5/00).

3 THE STOCASTIC TRANSLATION MODEL

The usual statistical translation models can be classified as single-word based (SWB) alignment models. Models of this kind assume that an input word is generated by only one output word [2]. This assumption does not correspond to the nature of natural language; in some cases, we need to know a word group in order to obtain a correct translation.

Recently, a simple alternative to these models has been proposed, the phrase-based (PB) approach [3][7]. The principal innovation of the phrase-based alignment model is that it attempts to calculate the translation probabilities of word sequences (phrases) rather than of only single words. These methods explicitly learns the probability of a sequence of words in a source sentence ($\tilde{f}$) being translated as another sequence of words in the target sentence ($\tilde{e}$).

As can be seen in Figure 1, we join words that are translated together in a natural way. Another property of our translation model is

Figure 1. Equivalent phrases in a sentence in Spanish, Portuguese, Italian, French and English

Se requerirá	una acción	de la Comunidad	para la
É necessária	uma acção	por parte da Comunidade	para
Sarà necessaria	un'azione	della Comunità	per dare
Une action	est nécessaire	au niveau communautaire	afin de
Action	is required	by the Community	in order to

that the alignment between pairs of phrase sequences is monotone-constrained. In the example, the first three sentences are monotone-translated. The generative process, which allows for the translation of a sentence in this model, can be broken down into the following steps: First, the input sentence is segmented into phrases. Then, each phrase is translated to the corresponding target phrase. The output sentence is built by concatenating the target phrases in the same order as in the source phrases.

To define de PB model, we segment the source sentence $\mathbf{f}$ into K phrases ($\tilde{f}_1^K$) and the target sentence $\mathbf{e}$ into K phrases ($\tilde{e}_1^K$). A uniform probability distribution over all possible segmentation is assumed ($\alpha(e)$).

$$Pr(\mathbf{f}|\mathbf{e}) = \alpha(e) \sum_K \sum_{\tilde{e}_1^K : \tilde{e}_1^K = \mathbf{e}} \sum_{\tilde{f}_1^K : \tilde{f}_1^K = \mathbf{f}} Pr(\tilde{f}_1^K | \tilde{e}_1^K) \qquad (2)$$

We assume a monotonous alignment, that is, the target phrase in position k is produced only by the source phrase in the same position:

$$Pr(\tilde{f}_1^K | \tilde{e}_1^K) = \prod_{k=1}^{K} p(\tilde{f}_k | \tilde{e}_k) \qquad (3)$$

where the parameter $p(\tilde{f}|\tilde{e})$ estimates the probability that the phrase, $\tilde{e}$, be translated to the phrase $\tilde{f}$. These are the only parameters of this model. A phrase can be comprised by a single word. Thus, the conventional word to word statistical dictionary is included. To estimate these parameters, we used a maximum verisimilitude approach using the EM algorithm [4]. One shortcoming of the PB alignment models is the generalization capability. If a sequence of words has not been seen in training, the model cannot reorder it properly.

From Eq. (1), the (sub)optimal target sentence $\mathbf{e}$ can be computed using a search algorithm. The design of such algorithm is a crucial part in statistical machine translation. Its performance directly affects the quality and efficiency of the translation. We use a search algorithm which is based on multi-stack-decoding [1]. This algorithm is very fast. We translated more than a hundred words per second in the experiments described in this paper.

4 EVALUATION

In order to carry out our evaluation, we have translated 120 sentences (2,456 words) using 4 different Spanish-Catalan MTs. These sentences have been taken from different media: a newspaper, a technical manual, legal text, etc.

To evaluate the proposal presented in this paper, we use *word error rate* (WER), percentage of words, which are to be inserted, deleted or replaced in the translation in order to obtain the reference sentence [6]. WER can be obtained automatically by using the editing distance between both sentences.

Our statistical translator is accessible at `http://ttt.gan.upv.es/~jtomas/trad`. Its components have been inferred automatically from training pairs using statistical methods [3]. For this propose, we used the corpora obtained in section 2. The results of our experiment can be observed in Table 1.

Table 1. WER obtained for 4 Spanish-Catalan translators. Taval-http://www.cultgva.es", Incita-http://www.incyta.com", Internostrum-http://www.internostrum.com.

Salt	Statistical	Incita	Internostrum
9.9	10.7	10.9	11.9

5 CONCLUSION

A system for automatic translation between the Spanish and Catalan languages has been presented. With the exception of fragmentation, all components were inferred automatically from training pairs. For the language model, we used a conventional trigram model. For the translation model, we selected a phrase-based model. A Maximum Likelihood Estimation criteria was used for training the models. A monotone multi-stack-decoding algorithm was used for searching.

We estimated the parameters of our models using the "DOGV" and the "El Periódico" corpora. The process of extracting and aligning these corpora has been presented in this work. Finally, we have presented the performance results of the system, and compared with other Spanish-Catalan translators. We have obtained translation results similar to the other commercial systems.

By using, a inductive approach which only requires a minimum amount of human intervention, we have obtained results which are similar to those obtained by the costly, knowledge-based systems. In these systems, linguistic experts attempt to transfer the knowledge that is necessary to resolve the problem.

ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish project TIC2003-08681-C02-02 the IST Programme of the European Union under grant IST-2001-32091.

REFERENCES

[1] A. L. Berger, P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. R. Gillett, A. S. Kehler, and R. L. Mercer. Language translation apparatus and method of using context-based translation models. United States Patent, No. 5510981, April 1996.

[2] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, 'The mathematics of statistical machine translation: Parameter estimation', *Computational Linguistics*, **19**(2), 263–311, (1993).

[3] J. Tomás and F. Casacuberta, 'Monotone statistical translation using word groups', in *Procs. of the Machine Translation Summit VIII*, pp. 357–361, Santiago de Compostela, Spain, (2001).

[4] J. Tomás and F. Casacuberta, 'Combining phrase-based and template-based models in statistical machine translation', in *Pattern Recognition and Image Analisys*, eds., F.J. Perales, A.J.C. Campillo, N. Pérez de la Blanca, and A. Sanfeliu, volume 2652 of *Lecture Notes in Computer Science*, 1021–1031, Springer-Verlag, (2003). 1st Iberian Conference, IbPRIA-2003.

[5] J. Tomás, F. Fabregat, J.M. del Val, F. Casacuberta, D. Picó, A. Sanchís, and E. Vidal, 'Automatic development of spanish-catalan corpora for machine translation', in *Procs. of the Second International Workshop on Spanish Language Processing and Language Technologies*, Jaén, Spain, (2001).

[6] E. Vidal, 'Finite-state speech-to-speech translation', in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pp. 111–114, Munich, Germany, (April 1997).

[7] R. Zens, F. J. Och, and H. Ney, 'Phrase-based statistical machine translation', in *Confer-ence on Empirical Methods for Natural Language Processing*, (2002).