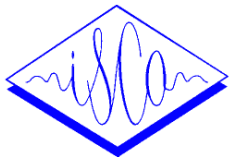

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

USING X-GRAMS FOR SPEECH-TO-SPEECH TRANSLATION

Adrià de Gispert *José B. Mariño*

TALP Research Center
Universitat Politècnica de Catalunya
{agispert|canton}@talp.upc.es

ABSTRACT

In this paper, a statistical speech-to-speech translation system, developed at TALP during the last months, is presented. By adapting well-known speech recognition techniques to the specific translation setting, the system is able to integrate speech signal into a finite state transducer that translates statistically domain-constrained Spanish sentences into English ones.

1. INTRODUCTION

Statistical machine translation is based on the assumption that every sentence **t** in the target language is a possible translation of a given sentence **s** in the source language. The main difference between two possible translations of a given sentence is a probability assigned to each, which is to be learned from a bilingual text corpus. Solving the translation problem is then finding the sentence in the target language that maximises equation (1).

$$\hat{t} = \arg\max_{t} \Pr(t \mid s) = \arg\max_{t} \Pr(s)\Pr(s \mid t) \quad (1)$$

Given the enormous variability of spoken language and the scarce availability of parallel corpora, an important question is how to build systems that are flexible enough to allow for several input alternatives and robust enough to provide a reasonable translation, from a relatively small training corpus.

When dealing directly with the speech signal, the objective is finding the sentence t in the target language that maximises the following equation:

$$\hat{t} = \arg\max_{t} \max_{s} \Pr(s,t)\Pr(x \mid s) \quad (2)$$

where x is the acoustic input signal and s is its decoding in the source language, doing the only assumption that the acoustic signal x does not depend on the target-language sentence t.

In this case, translation can be viewed either as a two-step process comprising a non error-free speech recognition phase and a subsequent translation phase that looks for a reasonable translation of the recognised text [10], or as an integrated recognition-translation process.

The first approach, also referred to as "serial or sequential" architecture, uses the decomposition of equation (3) to split the problem in two: a decoding stage (equation 4) and a translation stage (equation 5),

$$\hat{t} = \arg\max_{t} \max_{s} \Pr(s)\Pr(x \mid s)\Pr(t \mid s) \quad (3)$$

$$\begin{cases} \hat{s} = \arg\max_{s} \Pr(s)\Pr(x \mid s) & (4) \\ \hat{t} = \arg\max_{t} \Pr(t \mid \hat{s}) & (5) \end{cases}$$

On the other hand, the second approach, referred to as "integrated architecture" and followed by [4], tries to estimate directly the probability product in equation (1), following for example an X-gram approach. This alternative is presented in this paper, describing X-grams through a finite-state transducer that introduces acoustic models to integrate both processes in a single search.

The organization of the paper is as follows. In Section 2 the finite-state translating transducer is presented. In Section 3 the complex training of the system is discussed, introducing alternative approaches. Section 4 explains a first experiment on translation from Spanish to English carried out at TALP during the last months, whereas Section 5 draws conclusions and outlines further research that is to be done in the near future to improve the translation quality.

2. THE TRANSLATION SYSTEM

The translation process is done by means of a finite-state transducer whose edges are labelled with an extended symbol or bi-language unit. That is, each edge has a label that relates one word in the source-language to zero, one or more words in the target language.

Once the transducer is built, all well-known decoding techniques can be used to find the best-scoring translation of a given sentence. Viterbi and beam search can be used forwards only considering words in the source language (first word of the bi-language unit), whereas words in the target language are read during trace-back, as expressed in Figure 1.

To model this bi-language, non-deterministic X-grams [2] are especially advisable, since they estimate N-grams by means of a finite-state automaton considering variable memory lengths, thus taking into account long expressions that repeat itself in domain-constrained spoken language ("*me podría decir dónde*", "*I would like to know*",…), while achieving very efficient recognition (in this case translation) results by reducing perplexity.

This research was supported by the European Union under grant IST-2000-28323 (FAME project).

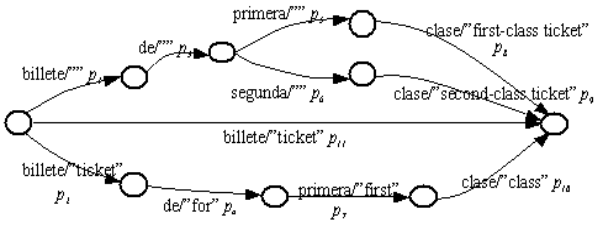

Figure 1: A translator FST from Spanish to English.

Apart from that, its layer architecture allows an easy introduction of the phonetic transcriptions and hidden Markov models of the source-language words to transform the text-to-text translator in an integrated speech translator.

3. TRAINING

Statistical machine translation depends entirely on the training corpus used to build the finite-state transducer. This learns the probabilities that a given bi-language unit follows or is preceded by any others through examples. Then, the process of creating the bi-language units that will train the transducer, deciding what word in the source language is connected to what words in the target language, is critical to the translation process. The three basic stages previous to the training are presented in the flow diagram of Figure 2.

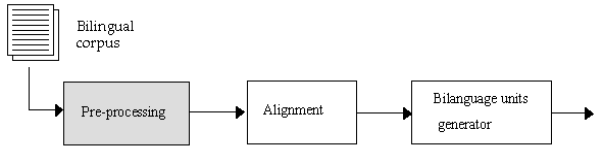

Figure 2: The three stages previous to training the translation transducer

The result of this process is a set of bi-language units, whose Language Model (or bi-language model) is to be learned by X-gram.

3.1. Pre-process of the bilingual corpus

The pre-processing stage is aimed at categorising words in order to reduce output vocabulary, helping the alignment stage to increase accuracy, without reducing input flexibility. Some basic word groups have been categorised, namely personal names, names of cities, towns or countries (manually), and dates, times of the day and numbers (automatically).

With the X-gram software, these categories or word groups can be easily modelled by smaller finite-state transducers that translate each of their possible alternative values.

3.2. Text alignment

The first stage when aligning two parallel texts is sentence alignment, that is, deciding which sentences in one language are a translation of which in the other language. This can be done statistically using many techniques, but a prevalent one is modelling sentence length as in [6]. As about word alignment, that is, alignment between elements inside the sentence, the "EGYPT" software (particularly the "giza" program) is used, which implements well-known IBM models [3] up to a simplified model 4, also called 4', as implemented by [7]. Some research on evaluating and improving these alignment models has already been presented [11].

This powerful alignment tool links zero (NULL word), one or more target-language words to each source-language word, considering all possible word positions in a sentence. However, this has a limitation, because only null-to-many, one-to-many and one-to-null alignments are allowed. When finding expressions that would need a many-to-many alignment (such as "`por favor`" and "`please`" in Figure 3), the alignment introduces unnecessary NULL words at the target language side (the word "`por`" is aligned to no English word, producing a bi-language unit like "`por/NULL`"), thus weakening the automat's capacity to learn from past alignments.

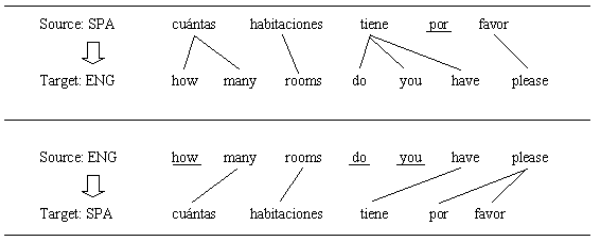

Figure 3: One-to-many alignments shorten the memory of the translation transducer.

To avoid this behaviour, the experiment can be repeated changing the source and target languages, and combining both results. Many combination strategies may be followed [10], but our simple approach consists of minimising the number of NULL words when both alignment directions do not produce same solutions. In this case, the combination of both alternatives should produce the following units:

```
cuántas/how_many
habitaciones/rooms
tiene/do_you_have
por_favor/please
```

3.3. The generation of bi-language units

Since the alignment tool considers all word position in a sentence without restrictions, it usually produces crossed alignments, as in Figure 4. However, the bi-language units generator has to build units so that the order of the sentence in both languages is not violated, a necessary requirement when dealing with finite-state translation transducers, as already exposed in [5], because otherwise the transducer would learn order-incorrect sentences.

A possible solution to this problem is to align the output word to the first input word that does not violate the order [5]. This approach, illustrated in Figure 4, allows the transducer to learn order-correct sentences, but it has two problems. On the one hand, it also shortens the automat's memory by introducing NULL or empty words (either in the source or the target side of the bi-language unit), as shown in solutions a) and b). This can be solved by joining all words in a single unit, as in c).

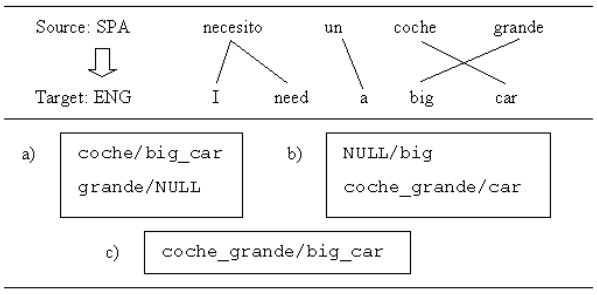

Figure 4: Alternative bi-language units when facing a crossed alignment.

Unfortunately, in both cases the independence of each alignment pair is broken during the joining, involving an information loss. This is a huge problem for instance when dealing with adjectives in English and Spanish, since they tend to precede or follow the noun, respectively, and then a different bi-language unit is needed to train the same noun combined with all possible adjectives.

Other approaches to this may be followed, such as introducing some kind of reordering in the target language sentence [1]. This would ideally allow to train a single unit for each adjective and a single unit for each noun without loss of flexibility, but the feasibility of such an approach is yet to be tested.

4. A FIRST EXPERIMENT

A first translation experiment from Spanish to English was carried out to get in touch with the practical difficulties when building such a translation system. The following sections deal with the parallel corpus used to train the translator, the training process itself and two translation experiments, one using text as input and one using speech signal.

4.1. Parallel corpus

Due to availability reasons, a small parallel corpus of 6,200 sentences in English and Spanish (3,100 sentences each, amounting to ~18.700 and 16.500 words, respectively) dealing with the tourist domain (mainly transportation, lodging, commerce and entertainment issues), was collected manually. The perplexity of the Spanish sentences was 33 (53 without punctuation marks, as in spoken language), and of the English ones was 28 (40 without punctuation). Unfortunately, up to 52% of the 2,650 different Spanish words appeared only once in the corpus (46% of 2,044 in English) due to the reduced size of the parallel corpus.

4.2. Training

When categorising, a word or group of words is substituted for its category name in both languages. During the training we expect that the categories in both languages will be matched to each other in the same bi-language unit, because we want to learn the upper-lever alignment (categories alignment in the form "`@PERSONA/@PERSON`") so that we can travel through the lower-level (any personal name) during decoding and output the correct translation. However, this may not be the case, because the alignment software may return a crossed matching and the generator will break the link between a category in the source language and its equivalent in the target language to preserve sentence order. To solve this problem, we have forced categories in both languages to be always linked in the same bi-language unit. This way, if we have the following example,

```
Source: habitación de @PERSONA
Target: @PERSON 's room
```

and "`@PERSONA`" is aligned to "`@PERSON`" whereas "`habitación`" is aligned to "`room`", the bi-language unit will be:

```
habitación_de_@PERSONA/@PERSON_'s_room
```

This bi-language unit apparently requires a finite-state transducer with all the possible personal names preceded by "`habitación de`" as input, or in general, it seems that it is necessary to replicate the transducers of all the variables whenever a bi-language unit is found that is different from "`S_CTG/T_CTG`", `S_CTG` and `T_CTG` meaning source and target category name. However, the layer architecture of the X-gram implementation offers a good solution to this inefficiency, because once the upper-level (bi-language units level) path is output, it is capable of searching through the lower level (word level) to look for the specific input category representation that lies under the category name. A list can offer then its translation. This way, only one finite-state transducer per category is needed, no matter the bi-language unit in which this category is found, as long as it appears in the source and the target sides.

As about alignment, as the corpus was already split into sentences, no paragraph or sentence alignment between both languages was necessary. The number of bi-language units created using the information of the alignment software was ~17.500, with 5,314 different units. Up to 72% of the units appeared only once in the training. When combining the two alignment directions, the number of bi-language units decreased slightly (~17.000), and the number of different units increased slightly (5.407), what draws a worse scenario because 74% of the units appear only once. These problems are to be overcome as soon as more training sentences are available. It will be then more advisable to compare and evaluate both approaches (without or with combination).

A limitation of the current implementation of X-grams (inherited from the speech recognition framework) is that it does not allow for NULL words to begin a sentence, because it makes no sense to begin a speech recognition process detecting silences. As it was discussed above, the alignment produces also null-to-many matches, thus generating bi-language units like "`NULL/word`" that might begin a sentence. A temporary solution to this has been to prevent all bi-language units from having a NULL in the source language side by linking this units to the first following one with a word different to NULL.

To overcome this problem, a modification of the X-gram implementation is planned for the near future.

4.3. Text-to-text translation

A test set of 100 sentences in Spanish was collected and used as input to the translator. The results were not promising in terms of sentence comparison, because only 20 sentences coincided entirely to the human translation, but this is not an

adequate measure when comparing translation solutions. Word error rate was 31,5%, but again this is too pessimistic a measure due to the fact that there is not a single correct translation to each input sentence. Actually, many cases like "do you have" in the reference and "have you got" in the system's output were found in the test, adding inexistent errors to the measure. This effect can be minimised if there exists a translation criteria when preparing the training corpus and the reference files so that the minimum number of target-language structures is used to convey the message of all the sentences. In practice, that means unification between synonyms in the target language.

4.4. Speech-to-speech translation

As about speech translation, the same set of 100 sentences was recorded three times on the phone by 20 speakers (15 utterances each), sampled at 8kHz and quantified using the A-law at 8 bits per sample. The phonetic representation unit is the demiphone [8], obtained through clustering as explained in [9]. The recognition models are 750 units, trained with 25 hours of Spanish speech obtained from the SpeechDat database [12].

Unfortunately, at deadline time the speech translation experiment has not been finished yet. Its results will be reported at the conference.

5. CONCLUSIONS

The TALP speech translation system has been introduced. Like the TALP speech recognition system, the new system is based on X-gram language modelling. Now, the X-gram automata provides the framework for a finite-state translator. Up to the present experience, the X-gram approach seems to be flexible enough for the task. However, its current ability can not be definitively assessed with the current lack of training and test material.

New parallel sentences in the same domain are already being collected, and new results with larger training and test corpora are to be reported in the near future.

5.1. Further research

One of the biggest disadvantages of the approach to translation through finite-state transducers is the sentence order limitation when creating the bi-language units. This limitation makes the system domain-constrained because the number of bi-language units grows exponentially as explained above. Ways of overcoming this limit are to be explored, either improving categorization rules to widen its effect on the translation results or including some kind of order mark in the bi-language unit, such as:

```
nombre/noun
+1/-1
adjetivo/adjective
```

reflecting that the English adjective should go one position before the noun (-1) in the target language (English), and one position after the noun (+1) in the source language (Spanish).

The layer structure of the X-gram implementation is also to be explored as a source of solving these cases, allowing for example the typical cross between noun and adjective in Spanish and English to be preserved, avoiding the presented loss of information.

As already mentioned, the limitation that there cannot be NULL words in the source language size of the bi-language unit due to the X-gram current implementation is to be overcome, and its effect, if any, in the translation is to be reported.

And last but not least, new evaluation strategies are to be researched, including not only unification of the structure in the target language as mentioned above, but also experimenting with semantic error measures, either manual or automatic.

6. ACKNOWLEDGEMENTS

The authors want to thank Antonio Bonafonte for his help in implementing many of the ideas presented in this paper.

7. REFERENCES

[1] Bangalore, S. and Riccardi, G. *A Finite-State Approach to Machine Translation,* North American ACL 2001 (NAACL-2001) , Pittsburgh, May 2001.

[2] Bonafonte, A. and Mariño, J.B. *Language modeling using X-grams.* Proc. ICSLP'96, pp. 394-397, Philadelphia, USA, October 1996.

[3] Brown, P., Pietra, S. D., Pietra, V. D. and Mercer, R. *The mathematics of statistical machine translation*. Computational Linguistics, 19(2), 1993.

[4] Casacuberta, F. *Finite-state transducers for speech input translation*. IEEE Automatic Speech Recognition and Understanding Workshop, Madonna di Campiglio, Italy, Dec. 2001.

[5] Casacuberta, F. *Inference of finite-state transducers by using regular grammars and morphisms*. Lecture Notes in Computer Science Vol. 1891, 2000.

[6] Gale, W. and Church, K.W. *A program for aligning sentences in bilingual corpora*. Proc. of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, 1991, pp.177-184.

[7] Knight, K., Al-Onaizan, Y., Purdy, D., Curin, J., Jahr, M., Lafferty, J. Melamed, D., Smith, N., Och, F.J. and Yarowsky, D. *EGYPT software*. JHU Workshop 1999. Available at http://www.clsp.jhu.edu/ws99/projects/mt/

[8] Mariño, J.B., Nogueiras, A., Pachès, P. and Bonafonte, A. *The demiphone: an efficient contextual subword unit for continuous speech recognition*. Speech Communication, Vol. 32, No. 3, pp. 187-197 (Oct. 2000).

[9] Mariño, J.B. and Nogueiras, A. *Top-down bottom-up hybrid clustering algorithm for acoustic-phonetic modeling of speech*. Proc. EUROSPEECH'99, pp. 1343-1346, Budapest, Hungary (Sept 1999)

[10] Ney, H., Nießen, S., Och, F.J., Sawaf, H., Tillmann, C. and Vogel, S. *Algorithms for statistical translation of spoken language*. IEEE Trans. on Speech and Audio Processing Vol. 8, No. 1, pp. 24-36, Jan. 2000.

[11] Och, F.J., Tillmann, C. and Ney, H. *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.

[12] http://www.speechdat.org