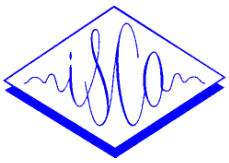
ISCA Archive
http://www.isca-speech.org/archive



4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

Learning Language Translation in Limited Domains using Finite-State Models: some Extensions and Improvements*

J. M. Vilar†
Depto. de Sist. Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera s/n
46071 Valencia (Spain)
E-mail: jvilar@dsic.upv.es

A. Marzal
Depto. de Informática
Campus de Penyeta Roja
Universitat Jaume I
12071 Castelló (Spain)
E-mail: amarzal@inf.uji.es

E. Vidal
Depto. de Sist. Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera s/n
46071 Valencia (Spain)
E-mail: evidal@dsic.upv.es

10.21437/Eurospeech.1995-307

Abstract

The Onward Subsequential Transducer Inference Algorithm (OSTIA) has been used for learning Language Translations in limited domain tasks. Although it is known to converge to the correct model when presented with enough training examples, the amount of training data can be prohibitive for large vocabularies. We address this problem by using appropriate clustering of words in both the input and output languages. Experimental results are presented which show that this approach effectively avoids dependency on the size of the vocabulary.

1 Introduction

Recently, new techniques have been proposed to automatically learn Language Translation (LT) models from pairs of sentences of the input and output languages [1]. The model adopted is a finite-state transduction automaton known as Subsequential Transducer (SST) [2]. These techniques have proved quite successful for a number of artificial tasks [1] and also a more natural, limited-domain task recently proposed by Feldman et al [3]. Moreover, owing to the finite-state nature of (stochastic) SSTs, these models lend themselves quite appropriate for their integration with standard acoustic-phonetic models of the input language, thus yielding simple and effective *speech-input* LT systems, which are fully trainable from data of the task under consideration [4].

A Subsequential Transducer (SST) is a deterministic finite-state network that accepts sentences from a given input language and produces associated sentences of an output language. Each edge of the network has associated an input symbol and an output string. Every time an input symbol is accepted, the corresponding string is output and a new state is reached. After the whole input is processed, additional output may be produced from the last state reached in the analysis of the input [2].

Given a set of training pairs of sentences from a translation task, the *Onward Subsequential Transducer Inference Algorithm* (OSTIA) learns an SST that generalizes the training set [1]. The algorithm builds a straightforward prefix-tree representation of all the training pairs and moves the output strings towards the root of this tree as much as possible, leading to an "onward" tree representation. Finally a state merging process is carried out. The algorithm guarantees identification of the target transduction in the limit; that is, if the unknown target translation can be assumed to exhibit a subsequential structure, convergence to it is guaranteed whenever the set of training samples is representative or, simply, large enough [1].

Additionally, if models for the input and/or output languages are known, a recently introduced extended version of OSTIA can be used, which produces SSTs that only accept input sentences and only produce output sentences compatible with these models [4, 5].

SSTs base their translation ability on "delaying" the production of output words until enough of the input sentence has been seen to guarantee a correct output. For instance, the input Spanish sentence (from Feldman's task [3]), "`un triángulo mediano y claro está ...`" is translated into English as "`a medium light triangle is...`" by following a sequence of states in the SST such that the word "`un`" is translated as "`a`", the words "`triángulo`", "`mediano`" and "`y`" are translated as the empty string, the word "`claro`" is translated as "`medium light triangle`", and the word "`está`" is translated as "`is`".

Every word sequence whose translation must be delayed is "stored" by means of the states of the SST. When the number of (functionally equivalent) words increases, the required number of states can grow as much as $O(n^k)$, where n is the number of words and k the length of the sequence. Clearly, for realistic tasks, the amount of training data required to show all the possible combinations would be far beyond practical limits. Experiments presented in Section 6 show the impact of lexicon growth in both the number of states of the SST and the size of the training set required to learn a translation task.

On the other hand, it seems appropriate that no explicit training be required as new words are being incorporated to the task vocabularies. Our proposal addresses both objectives: to reduce impact of vocab-

* Work partially supported by Spanish CICYT under contract TIC95-0984-C02.
† Supported by a grant of the Spanish *Ministerio de Educación y Ciencia*

ulary growth and to avoid retraining the system every time the vocabulary is extended.

2 Experimental task

Experiments are presented on an extension of MLA, a pseudo-natural task recently proposed by Feldman et al [3]. The original task consisted of descriptions of simple two-dimensional visual scenes involving a few geometric objects with different shape, shade and size, and located in different relative positions. The original language of this task was extended to cover the possibility of adding or removing objects to or from a scene, and the task was adapted for LT experimentation [6].

The following examples illustrate typical sentences of this (extended) translation task:

- `a large ellipse is added far below the large light triangle` $\Longleftrightarrow$ `se añade una elipse grande muy por debajo del triángulo grande y claro`
- `the large dark triangle which is to the left of the small light circle and the dark square is removed` $\Longleftrightarrow$ `se elimina el triángulo grande y oscuro que está a la izquierda del circulo pequeño y claro y del cuadrado oscuro`

In order to study the impact of vocabulary growth in OSTIA, four extensions have been defined: EX0 with 6 shapes, 3 sizes, 2 shades; EX1 with 12 shapes, 5 sizes, 4 shades/colors; EX2 with 18 shapes, 7 sizes, 6 shades/colors; and EX3 with 118 shapes, 57 sizes, 56 shades/colors. The different shapes have been selected so that half of them are masculine and half feminine when translated into Spanish. Also, the different adjectives present their masculine and feminine forms.

The number of words of each sentence ranges from 2 to 28 with an average of 10.

3 Using word categories for translation

Many of the examples appearing on the training set represent different instances of similar patterns. For example, the pairs

[un cuadrado rojo toca una elipse
a red square touches an ellipse]

[un triángulo oscuro toca una línea
a dark triangle touches a line]

can be considered as instances of the pattern

[un NOUN ADJ toca una NOUN
a ADJ NOUN touches a NOUN]

Taking this into account, the translation process can be seen as a three stage procedure:

1. Words of the input sentence are substituted by appropriate category labels.[1]
2. The labeled input sentence is translated into a labeled output sentence by an automatically learned SST.
3. The labels of the output sentence are replaced by their respective translations.

In principle, the substitution of a word by its category label can be done in an straightforward way in limited domain tasks. For example, in the above described task the labeling algorithm simply replaces each word by its corresponding (unique) category label without taking into account any context information.

For the translation task used in our experiments, we have considered three categories: NOUN for shapes (nouns), ADJ for sizes and shades/colors (adjectives), ADV for adverbs.

In the first sentence of the above example, the labeling procedure would yield the following result:

```
un triángulo oscuro toca una línea
 ↓      ↓      ↓     ↓   ↓    ↓
un    NOUN    ADJ   toca una NOUN
```

Next, the labeled sentence would be considered as input to an automatically learned SST, producing the following translation:

```
un NOUN ADJ   toca una NOUN
 ↓   ↘ ↙     ↓    ↓   ↓
 a  ADJ NOUN touches a NOUN
```

Finally, every label in this sentence must be replaced by the translation of the word it substituted in the first stage. For instance, the label ADJ must be replaced by the translation of "oscuro", which is "dark". However, a problem arises with the replacement of the two instances of NOUN: there is no direct way to know which one corresponds to the word "triángulo" and which one to "línea". Some additional information must be kept along with the labels in order to solve this ambiguity. We have enriched labels with an attribute indicating its relative position in the sentence with respect to other words with the same label. The whole procedure can be depicted as follows:

```
un triángulo  oscuro   toca  una línea
 ↓     ↓       ↓      ↓     ↓    ↓
un   NOUN1    ADJ1    toca  una NOUN2
 ↓     ↘ ↙     ↓      ↓     ↓    ↓
 a    ADJ1    NOUN1  touches a NOUN2
 ↓     ↓       ↓      ↓     ↓    ↓
 a    dark   triangle touches a line
```

[1] In principle, every word might be substituted by a label, but only those belonging to categories with many members or whose number of members is expected to grow, *need* to be categorized.

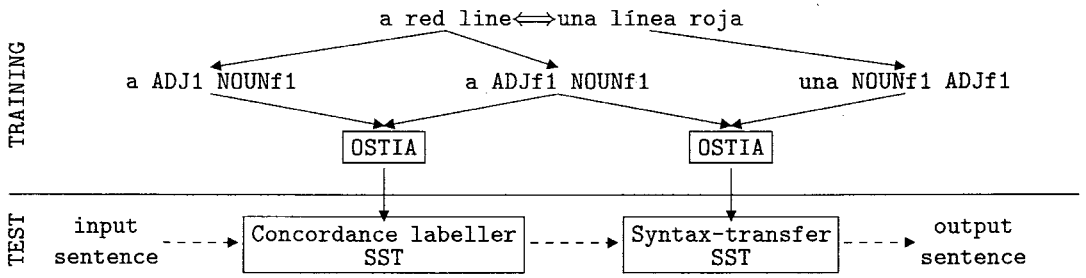


Figure 1: Schematic representation of the proposed methodology. From the pairs of the original corpus, two new corpora are obtained: one to train a word labeller SST and the other to train the syntax transfer SST. The input sentences are translated first with the word labeller and then with the syntax transfer SST.

When trying to apply the same three stage procedure to English-to-Spanish translation, another problem arises due to the fact that Spanish adjectives must agree in gender with nouns. For instance, if an ADJ label corresponding to the English word "dark" is to be replaced by its Spanish translation, it is not known whether "oscuro" (masculine) or "oscura" (feminine) should be used, unless context information is available (gender of the corresponding noun).

In general, this problem arises whenever an input language word has several possible translations depending on inflections in the output language (eg Spanish, German, etc). We have addressed this problem by using categories for the output language that carry this information and by using these categories to label the input sentence. In the following example m is used to indicate masculine and f feminine:

a	dark	triangle	touches	a	line
a	ADJm1	NOUNm1	touches	a	NOUNf2
un	NOUNm1	ADJm1	toca	una	NOUNf2
un	triángulo	oscuro	toca	una	línea

Now the first stage of translation needs a more elaborate categorizer and, in fact, we have considered it as a special case of translation. An SST learned by OSTIA has been used to model this translation. This strategy allows us to follow the corpus based approach in the development of both the first and second stages, as well as to easily integrate these stages into a single SST by composition of the two previous SSTs.

4 Training the concordance labeller SST

The purpose of this SST is to perform categorizations like the following:

```
[ a dark triangle touches a line
  a ADJm1 NOUNm1 touches a NOUNf2 ]
```

Note that *knowing the translation of the input sentence* provides enough information to do this labeling. Therefore, it is possible to obtain a new corpus from the original one and to use it to train this SST with OSTIA.

In order to allow an easy extension of the vocabulary we divide the input language words in three different groups: a) words that need not be categorized, b) words with unique translation, c) words with ambiguous translation. In the first group we include function words (whose number will not increase as the lexicon grows) and, for Feldman's translation task, verbs (whose number can also be expected to remain constant).

An initial labeling of the input sentence can be made in a word by word basis: words in group a) are not replaced; words in group b) are replaced by their respective categories with disambiguation information[2]; words in group c) are replaced by their category without disambiguation information. In our example "a dark triangle touches a line" is labeled as "a ADJ1 NOUNm1 touches a NOUNf1".

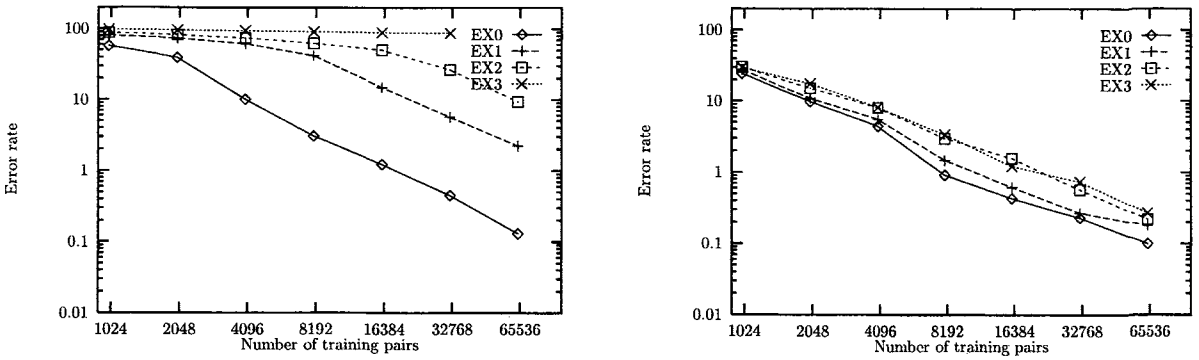
OSTIA is presented with a corpus consisting of pairs in which the first sentence has been labeled according to this scheme and the second one has been fully labeled (see Figure 1, left). The resulting SST is then adapted to the lexicon of the task by replacing each edge that has an input label in groups b) and c) by a set of edges, each one having as input one of the words in the group and the same output as the original edge.

5 Training the syntax-transfer SST

The purpose of the second SST is to perform translations like the following:

```
[ a ADJm1 NOUNm1 touches a NOUNf2
  un NOUNm1 ADJm1 toca un NOUNf2 ]
```

[2]This can always be done in a straightforward way by looking at its translation, eg **square** is always translated as **cuadrado**, which is a masculine word in Spanish.



(a) Conventional approach.

(b) Using word categories.

Figure 2: Whole sentence error rates of the SSTs for increasing training set size. Spanish/English vocabulary sizes: EX0: 37/28, EX1: 50/38, EX2: 63/48, EX3: 363/248.

In the previous section we have seen how the input sentences can be labeled. The labeling of the output sentences poses no problem, since they are labeled according to categories in the output language. The SST is trained using OSTIA from a corpus consisting of pairs like the one in the example. See Fig. 1, right.

6 Experimental results

Figure 2 displays the impact of vocabulary-size growth in terms of error rate as a function of the number of training pairs. The test set consisted of ten thousand sentences. Figure 2-(a) shows results for EX0, EX1, EX2, and EX3 tasks without word clustering. It can be seen that even small increases in vocabulary size dramatically increase the number of examples required to train the SST with OSTIA. Figures 2-(b) shows the error rate of the transducer obtained composing the concordance labeller SST and the syntax-transfer SST. In this case the dependency on the size of the vocabulary is negligible, as we expected.

On the other hand, the same experiments discussed above have been carried out for the translation from Spanish into English with better results than those here reported. This tasks is in fact significantly easier than English-into-Spanish translation since no gender concordance problems appear in this case. These results are omitted in this paper for the sake of brevity.

7 Integration with speech input

Due to the finite-state nature of the SSTs, they can be easily integrated with lexical and acoustic-phonetic models, allowing the use of traditional search methods such as the One Stage algorithm, Stack Decoding, etc. The output of the integrated translator is the sequence of output words of the edges in the optimal path for a given utterance, which can then be replaced by the corresponding words as explained above.

8 Concluding remarks

The work presented in this paper constitutes a first step towards enabling a scale-up of transducer learning techniques that have been previously shown quite successful for small speech-input language translation tasks [4]. We have used here manually derived word categorizations for both the input and output languages. The next step is trying to automatically obtain such categorization from the training corpora of the tasks. For this purpose, we plan to study suitable modifications of techniques such those presented in [7, 8]

References

[1] J. Oncina, P. García, E. Vidal: "Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks". *IEEE Trans. on PAMI*, Vol. 15, No. 5, pp. 448–458, 1993.

[2] J. Berstel: *Transductions and Context-Free Languages*. Teubner, Stuttgart. 1979.

[3] J.A. Feldman, G. Lakoff, A. Stolcke, S.H. Weber: "Miniature Language Acquisition: A touchstone for cognitive science". Technical Report TR-90-009. ICSI, Berkeley, CA, USA, 1990.

[4] V.M. Jiménez, A. Castellanos, E. Vidal, J. Oncina "Some Results with a Trainable Speech Translation and Understanding System". Proc. of ICASSP95.

[5] J. Oncina, A. Castellanos, E. Vidal, V.M. Jiménez: "Corpus-Based Machine Translation through Subsequential Transducers". Proc. of *3rd ICSNLP*, Dublin, Ireland, 1994.

[6] A. Castellanos, I. Galiano, E. Vidal: "Application of OSTIA to Machine Translation Tasks". In *LNAI (862): Grammatical Inference and Applications*. R.C. Carrasco & J. Oncina (eds.), pp. 93–105, Springer-Verlag, 1994.

[7] F. Jelinek, R.L. Mercer, and S. Roukos: "Classifying Words for Improved Statistical Language Models", *Proceedings of the ICASSP-90*, Albuquerque, NM, USA, pp. 621–624, 1990.

[8] R. Kneser and H. Ney: "Improved Clustering Techniques for Class-Based Statistical Language Modelling ", *Proceedings of the EUROSPEECH-93*, Berlin, Germany, pp. 973–976, 1993.