Spoken-Language Machine Translation in Limited Domains: Can it be Achieved by Finite-State Models? *

J. M. Vilar† (1) A. Castellanos† (1) V. M. Jiménez‡ (1)

J. Oncina (2) H. Rulot (3) J. A. Sánchez (1) E. Vidal (1)

(1) Depto. de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino de Vera s/n, 46071 Valencia (Spain)
E-mail: jvilar@dsic.upv.es
(2) D.T.I.C., Universidad de Alicante (Spain)
(3) Centro de Informática, Universidad de Valencia (Spain)

Abstract

Subsequential transducers constitute a formal model for translation that may be considered perhaps too simple to model translation between natural languages. However, their capability can suffice in limited-domain translation tasks. The finite-state nature of subsequential transducers makes their integration with well-known Continuous Speech Recognition technology both easy and efficient. A recent algorithm allows the automatic learning of these transducers, given a sufficiently large set of examples of sentences and their corresponding translations, and it also allows the incorporation of syntactic restrictions of the input and/or output languages. In this paper, we describe an implementation of a Speech Translation System for limited domains which is fully trainable and capable of real time translation from speech input.

1 Introduction

The problems of Machine Translation (MT) and Language Understanding (LU), when considered in their vast generality, are far from being satisfactorily solved. Interestingly, however, in contrast with such a *general* formulation, many MT and LU tasks of interest to industry and business have *limited domains;* that is, lexicons are of small size and the universe of discourse is limited: reservation of flights, hotels, etc.; tourist guide talks; broadcast of weather reports; etc.

Although natural languages are complex, the mappings defined by their translations can be comparatively much simpler, specially when these languages are close as is the case with many European languages. LU can also be seen as a particular case of translation;

* This work has been partially supported by Spanish C1CYT under contract TIC92-1026-C02.
† Supported by the Spanish *Ministerio de Educación y Ciencia*
‡ Supported by the Spanish *Conselleria d'Educació i Ciencia de la Generalitat Valenciana.*

that is a translation from a natural language into a formal one, in which we can adequately specify how the machine has to act in response to the input sentence. For instance, a natural language question to a database can be "translated" into the corresponding formal query, which can then be used to obtain the adequate answer.

When considering speech input operation new problems arise. Most of the current efforts to cope with this problem are based on the use of previously developed *text-input* LT or LU systems (generally relying on knowledge-based technology), which are serially coupled to the output of state-of-the-art *speech recognition* front-ends [12, 13, 17, 18]. Such a procedure is quite sensitive to front-end errors, since it does not exploit the powerful intrinsic restrictions that underlie the output language syntax and the translation rules, to conveniently guide the search at the (input) acoustic and lexical levels. A possibly better approach would be trying to solve the LT and HI problems under a framework closer to the standard assumptions under which successful speech front-ends are developed. This means devising adequate models for LT and LU which: i) can be *automatically learnt* from training data for each task considered; and ii) can be combined with the input-language acoustic and lexical models into an appropriate *integrated network,* in which an optimal search to find the best output can be performed [14]. In these situations, an adequate model is given by *subsequential transduction,* both for its simplicity and for the availability of an efficient algorithm for learning the corresponding finite-state devices.

2 Subsequential Transducer Learning

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 [2]. The translation of an input sentence is performed departing from an initial state and accepting the input symbols one by one. 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 accepted additional output may be produced from the last state reached. This final output differentiates SSTs from pure sequential transducers and allows them to overcome several limitations of the latter [14].

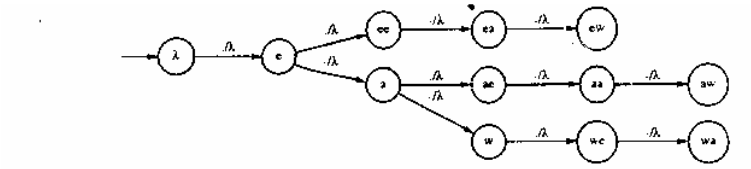
Given a set of training pairs of sentences from a translation task, the *Onward Subsequential Transducer Inference Algorithm* (OSTIA) efficiently learns a SST that generalises the training set [10]. Moreover, if the unknown target translation can be assumed to exhibit a *Subsequential* structure [2], convergence to this translation is guaranteed if the set of training samples is representative or, simply, large enough [10]. To illustrate the functioning of the algorithm, a small example for a simple task can be seen in Figure 1. The task is to decode Morse messages which can contain only three letters; a(• –), e(•) and w(• – –) but without any marks between letters. The algorithm is provided with the training pairs shown in Figure l(a), and proceeds in three stages:

1. The input sentences are first represented in a prefix tree. 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 node reached by the corresponding input string. Figure l(b) shows this initial tree for the above mentioned examples.

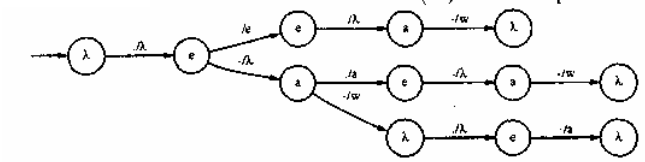
2. The longest common prefixes of the output strings are recursively moved, level by level, from the leaves of the tree towards the root. The onward prefix tree of the

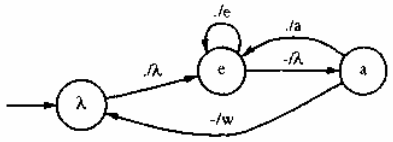
$\{(\lambda, \lambda), (\bullet, e), (\bullet\ \bullet, e\ e), (\bullet\ -, a), (\bullet\ \bullet\ -, e\ a), (\bullet\ -\ \bullet, a\ e), (\bullet\ -\ -, w), (\bullet\ \bullet\ -\ -, ew),$
$(\bullet\ -\ \bullet\ -, aa), (\bullet\ -\ -\ \bullet, we), (\bullet\ -\ \bullet\ -\ -, aw), (\bullet\ -\ -\ \bullet\ -, wa)\}$.

(a) Set of training samples.



(b) Initial prefix tree.





(c) Onward prefix tree. (d) Final subsequential transducer.

Figure 1: The three steps of OSTIA for a simple Morse decoder.

examples is presented in Figure l(c).

3. Starting from the root, 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. Figure l(d) shows the result of this process for the Morse example.

All these operations can be very efficiently implemented, yielding an extremely fast algorithm that can easily handle huge sets of training data. The algorithm does not try to enforce any restriction either in the input or in the output languages of the SST. The resulting transducer may entail some of these restrictions as a by-product of the necessity of enforcing the translation rules themselves (as is the case in the example shown in Figure l(d)). However, in general OSTIA tends to generalise as much as possible the training data resulting in transducers that are very permissive in relation to the input and output syntax. While this has no negative effect when new correct input sentences are submitted to translation, the implications can be very negative for erroneous input data. This particularly applies to translation of *input speech,* a task were the robustness of the transducer is specially important: it should be able to produce approximately-correct translations for approximately well-recognised sentences [9. 11].

A naive approach to this problem consists in using an explicit input language model to obtain syntactically-correct sentences during the recognition phase and then translating the best hypothesis by means of the transducer learned by OSTIA, in a *decoupled* way. Such an scheme has the disadvantage of not taking the syntactic restrictions underlying the transducer itself (and those of the output language) into account for a better guidance during the speech-recognition stage.

Instead of that, we have used OSTIA with Domain and Range (OSTIADR), a recently introduced extended version of OSTIA which uses syntactic restrictions of the input and/or output languages, expressed by finite-state models, to guide the merging of states and limit the possible over-generalisations from the training data [11]. This version produces SSTs that only accept input sentences and only produce output sentences compatible with input and/or output models. In addition, text-to-text experimental results

have shown that the new version produces highly accurate transducers using less training samples [11].

This extension of OSTIA consists in a simple modification of the third stage, so that states can be merged only when the prefixes of the input and output strings leading to them reach the same states in the input and output automata, respectively. To test this efficiently, a single preprocess is done after completion of the second stage. In this preprocess the nodes of the tree are labelled according to the states reached in the automata and the above mentioned condition becomes a simple comparison of labels.

3 Overview of the Recognition and Translation System

Traditionally, Continuous Speech Recognition (CSR) systems have been used to *translate* an acoustic signal into the word sequence most likely uttered by the speaker. The approach of many current CSR systems is to model the task as the composition of different mappings between intermediate stages, i.e. from the acoustic signal into phonemes, and so on into words and syntactically correct sentences. Spoken LT (and, in particular, LU in the sense mentioned above), can be seen as the addition of a new mapping, from the sentence uttered by the speaker into a sentence in a *different language.*

So far, the best results in CSR have been obtained *integrating* the different sources of knowledge employed at each of the levels into a global model. The recognition is then viewed as a search for the most probable and syntactically correct word sequence given the acoustic signal. This integration can be easily accomplished if (stochastic) finite-state modelling is adopted at each level. The resulting integrated model is then a graph through which an optimal path can be found by Dynamic Programming.

This approach can be extended if the mapping used for the translation is also described by a finite-state model, as is the case for SSTs. The phonetic and lexical models of the input language are then embedded in the resulting SST to give the final model to be used in the search for the optimal translation performed in the recognition phase.

This methodology has been used in the development of a fully trainable Speech Translation and Understanding System [9]. This system is based on conventional *Viterbi Beam Search* through a network which embeds phonetic, lexical, syntactic, and translation stochastic finite-state models. Phonetic models are discrete Hidden Markov Models (HMM). Lexical models describe words in terms of valid concatenations of phonemes. Translation and syntactic models are SSTs embedding stochastic (finite-state) language models describing sentences in terms of possible concatenations of words.

Summarising, the system can be seen as the integration of three levels:

Syntactic and translation level: A finite state network, learned with OSTIADR, that describes the syntax of the input and output languages and the rules for the translation in terms of the words of both languages.

Lexical level: each word is described by a set of finite state automata, with phonemes associated with the arcs.

Phonetic level: A set of finite stochastic networks (HMM), each one describing one of the phonemes of the input language.

Spanish:	un circulo grande y claro y un triangulo están debajo de un cuadrado mediano
English:	a large light circle and a triangle are below a medium square
German:	ein grosser weisser Kreis und ein Dreieck sind unter einem mittleren Viereck
Semantic:	La(x) & Li(x) & C(x) & T(y) & M(z) & S(z) & B(x;z) & B(y;z)

Spanish:	se añade un cuadrado grande y claro encima del circulo oscuro
English:	a large light square is added above the dark circle
German:	man hat ein grosses weisses Viereck über dem dunklen Kreis hinzugefügt
Semantic:	La(x) & Li(x) & S(x) & D(z) & C(z) & A(x;z) & Ad(x)

Figure 2: Two examples of translations of Spanish sentences for the experimental task, above for the original task, below for the extended task.

Each of these components can be automatically learned. The integration of these components can be seen as the substitution of the edges of the automata describing the words for the HMMs describing the corresponding phonemes. These "extended" word models are in turn used to expand the edges of the transducers, giving rise to the final composite model. Note that this integration is only *virtual,* and only those arcs included within the beam of the Beam Search are actually expanded. Thanks to this limited expansion huge networks can be efficiently handled.

4 Experimental Results

4.1 Visual Scenes Description Task

The system has been tested with an extension of a pseudo-natural task recently proposed by Feldman et al. [7]. 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 and LU experimentation [5, 6]. In the present work, Spanish has been chosen as the input language; the output can be English or German for LT, or a semantic description of the scene in terms of first-order logic formulae for LU. The vocabulary size was between 25 and 70 words, depending on the language. Examples of these input and output sentences are shown in Figure 2.

4.2 Training the acoustic, syntactic and translation models

For the experimental results reported below, standard phonetic Hidden Markov Models and conventional lexical models are used, details can be seen in [9]. The syntactic restrictions of the input and output languages have been modelled by stochastic *k-Testable Automata* (*k*-TA), which are equivalent to *k-Grams* [3, 4, 8, 15].

A set of 50100 input/output paired (text) sentences (for each of the 3 different output languages) was obtained using a semi-automatic procedure. This procedure was driven by a Syntax-Directed Translation Scheme [1] governed by English context-free grammars for the basic and extended MLA tasks, along with the associated grammars for Spanish and German [5]. From this set, 100 input/output sentences were randomly selected for speech-input testing purposes. The remaining 50000 pairs were used to automatically learn different *k*-TA *(k* = 2,3,4) for the input and output languages as well as different

	Spk 1	Spk 2	Spk 3	Spk 4	Avg.
(i)	7.0 %	13.1 %	18.0 %	22.8 %	15.2 %
(ii)	5.2 %	1.7 %	1.4%	2.7%	2.8 %

(a) Spanish-English models

	Spk 1	Spk 2	Spk 3	Spk 4	Avg.
(i)	8.7 %	14.3 %	12.0 %	25.7 %	17.7 %
(ii)	7.3 %	1.5 %	0.7 %	3.3%	3.2 %

(b) Spanish-German models

	Spk 1	Spk 2	Spk 3	Spk 4	Avg.
(i)	11.5 %	15.8 %	16.8 %	21.3%	16.4 %
(ii)	5.5 %	2.7 %	3.8 %	3.5%	3.9 %

(c) Spanish-Semantics models

Table 1: Results with speech input (translation word error rates): (i) Decoupled scheme: recognition guided by the 4-TA of the input language, and translation performed with the transducers learned by OSTIA; (ii) Integrated scheme: recognition and translation guided by the transducers learned by OSTIADR using both the 4-TA of the input and output languages.

SSTs. For comparison purposes, both the above described *integrated* system and a *de-coupled* system in which *speech recognition* is performed prior to *translation,* have been implemented. For the decoupled approach SSTs where learned with the original OSTIA (without input or output syntactic restrictions) and recognition was performed using *k*-TAs of the input language. For the integrated approach SSTs were learned with OSTIADR using the input and output *k*-TAs, and a stochastic extension of the transducers was carried out by estimating the transition probabilities from their frequencies of use for processing the sentences in the training-set.

4.3 Recognition and Translation Results

From the randomly selected test-set of 100 input/output pairs, each Spanish test sentence has been uttered by four speakers (one of them -speaker number 2 in Table 1- also participated in the training of the HMMs). The system outlined above has been used to analyse these utterances, using the same parameter for the recogniser (beam search thresholds) in all the experiments.

Table 1 presents the translation word error rates (including insertion, substitution and deletion errors) for the four speakers and their averages. In general, a great improvement of the results is observed when syntactic constraints are integrated in the learned transducers.

The size (number of arcs) of the integrated SSTs was typically less than five times the size of the corresponding *k*-TAs in the case of Spanish-to-English and Spanish-to-German models, and up to 30 times the size of the corresponding *k*-TAs in the case of Spanish-to-Semantics models. This is due to a larger semantic vocabulary as well as to the higher degree of "asynchrony" in the Spanish-to-Semantic translation. In fact, the semantic representation was specifically chosen for studying this effect. For instance, in Figure 2 the Spanish segment *"se añade"* corresponds to "*Ad(x)"* which appears at the very end of the semantic representation. In spite of this increment in size, Viterbi beam search recognition-and-translation time was always lower using the integrated transducers, and never greater than 0.4 times real-time in a HP-9000/735 workstation.

5 Concluding Remarks

General purpose automatic translation (or understanding) of spontaneous speech is far from being satisfactorily solved. However, many applications of interest can be limited to a small or medium-sized vocabulary, and they have a restricted semantic domain. For tasks of this kind, it seems perfectly feasible to build systems by means of a finite-state model, not only of the syntactic constraints of the languages involved, but also of the required translation mapping itself. Moreover, since all these models can be automatically learned from training data, the building of Speech Translation and Understanding Systems for these tasks can be done at low development costs.

An important bottleneck of this approach lies in the availability of large corpora of paired sentences in different languages. It is likely that these resources will be available in the near future, in the same way as corpora for training acoustic and language models exist today. Note also that inherent to OSTIA is its independence from alignments at any sub-sentence level. This implies that the building of these corpora does not need to take care of these costly alignments.

The ability of SSTs to defer translation until the necessary amount of input data has been seen is clearly an advantage, but it can lead to models that grow excessively when confronted with large vocabularies which include many different possibilities for words in similar syntactic categories. The problem is that SSTs use the states as "storage" for the information seen. A typical example is the translation of Spanish noun phrases into English, which normally involves a reversal of the order of adjectives. That means that there must be at least an state for each possible combination of adjectives. A way of solving this problem is the use of word categories. The process of translation can then be decomposed in three stages: first the successive words that appear in the input sentence are labelled and substituted for their categories together with a numerical label indicating their relative position in the sentence; then the transducer is employed to obtain a similarly labelled sentence in the output language; finally the numbered labels, with help of a dictionary, are employed to substitute the output categories for their corresponding words. The advantage of this approach is that the category transducer can be learned using significantly less training pairs and have lower error rates and smaller sizes than the ones learned with the original approach [16].

References

[1] A. AHO, J. ULLMAN: *The Theory of Parsing, Translation and Compiling,* 1, Prentice Hall, 1972

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

[3] G. BORDEL, I. TORRES, E. VIDAL: "Back-off Smoothing in a Syntactic approach to Language Modelling". Proc. of *ICSLP-94.* Japan, September 1994.

[4] G. BORDEL, I. TORRES, E. VIDAL: "QW1: A Method for Improved Smoothing in Language Modelling". Proc. of *ICASSP-95.* Detroit, May 1995.

[5] A. CASTELLANOS, I. GALIANO, E. VIDAL: "Application of OSTIA to Machine Translation Tasks". In *Lecture Notes in Artificial Intelligence (862): Grammatical Inference and Applications.* R.C. Carrasco and J. Oncina (eds.), pp. 93-105, Springer-Verlag, 1994.

[6] A. CASTELLANOS, E. VIDAL, J. ONCINA: "Language Understanding and Subsequential Transducer Learning". Proc. of ICGI, Colchester (England), 1993.

[7] J.A. FELDMAN, G. LAKOFF, A. STOLCKE, S.H. WEBER: "Miniature Language Acquisition: A touchstone for cognitive science". Technical Report TR-90-009. International Computer Science Institute, Berkeley, CA, USA, 1990.

[8] P. GARCIA, E. VIDAL: "Inference of K-testable languages in the strict sense and applications to syntactic pattern recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 12(9), pp. 920-925. 1990.

[9] V.M. JIMÉNEZ, E. VIDAL, J. ONCINA, A. CASTELLANOS, H. RULOT, J.A. SÁNCHEZ: "Spoken-Language Machine Translation in Limited-Domain Tasks", In *Proceedings in Artificial Intelligence: CRIM/FORWISS Workshop on Progress and Prospects of Speech Research and Technology,* II. Niemann, R. de Mori and G. Hanrieder (eds.), pp. 262 265. Infix, 1994.

[10] J. ONCINA. P. GARCÍA, E. VIDAL: "Learning Subsequeritial Transducers for Pattern Recognition Interpretation Tasks". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 15, No. 5, pp. 448-458, May 1993.

[11] J. ONCINA, A. CASTELLANOS, E. VIDAL. V.M. JIMÉNEZ: "Corpus-Based Machine Translation through Subsequential Transducers". Proc. of *3rd International Conference on the Cognitive Science of Natural Language Processing,* Dublin, Ireland, 1994.

[12] M. RAYNER, I. BRETAN, D. CARTER: "Spoken Language Translation with mid-90's Technology: A Case Study", Proc. of *EUROSPEECH-93,* Berlin (Germany), September 1993.

[13] D.B. ROE, F.C.N. PEREIRA, R.W. SPROAT, M.D. RILEY, P.J. MORENO, A. MACARRÓN: "Efficient Grammar Processing for a Spoken Language Translation System", Proc. of *ICASSP-92,* pp. 213 216, 1992.

[14] E. VIDAL: "Language Learning, Understanding and Translation" In *Proceedings in Artificial Intelligence: CRIM/FORWISS Workshop on Progress and Prospects of Speech Research and Technology,* H. Niemann, R. de Mori and G. Hanrieder (eds.), pp. 131-140. Infix, 1994.

[15] E. VIDAL, F. CASACUBERTA, P. GARCÍA: "Syntactic Learning Techniques for Language Modeling and Acoustic-Phonetic Decoding". In *Speech Recognition and Coding: New Advances and Trends,*]. Rubio and J.M. López (eds.), Springer-Verlag, 1994.

[16] J.M. VILAR, A. MARZAL, E. VIDAL: "Learning Language Translation in Limited Domains using Finite-State: some Extensions and Improvements". To appear in Proc. of EUROSPEECH-95.

[17] W. WAHLSTER: "Verbmobil: Translation of Face-to-Face Dialogs", Proc. of the *MT Summit IV,* Kobe, Japan, 1993.

[18] M. WOSZCZINA, N. AOKI-WAIBEL, F. D. BUO ET AL.: "JANUS 93: Towards Spontaneous Speech Translation", Proc. of *ICASSP-94,* Vol. 1, pp. 345-348. 1994.