SOME RESULTS WITH A TRAINABLE SPEECH TRANSLATION AND UNDERSTANDING SYSTEM [1]

V. M. Jiménez[2], *A. Castellanos*[3] *and E. Vidal*

Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Spain
E-mail: vjimenez@dsic.upv.es

ABSTRACT

The problems of Limited-domain Spoken Language Translation and Understanding are considered. A standard Continuous Speech Recognizer is extended for using automatically learnt finite-state transducers as translation models. Understanding is considered as a particular case of translation where the target language is a formal language. From the different approaches compared, the best results are obtained with a fully integrated approach, in which the input language acoustic and lexical models, and (N-gram) Language Models of input and output languages, are embedded into the learnt transducers. Optimal search through this global network obtains the best translation for a given input acoustic signal.

1. INTRODUCTION

As speech processing techniques become increasingly able to cope with many real-world Continuous-Speech Recognition (CSR) applications, more ambitious targets are being considered, such as speech-input Language Translation and Understanding (LT and LU). Both problems can be uniformly formulated if we assume that the ultimate goal of a LU system is to drive the actions associated to the meaning conveyed by the sentences issued by the users. In this case, the understanding problem simply becomes one of translating the natural language sentences into formal sentences of an adequate (computer) command language. For example, "understanding" natural language (spoken) queries to a database can be seen as translating these queries into an appropriate computer-language code to access the database. Clearly, under such an assumption, LU can be seen as a (possibly simpler) case of LT in which the output language is *formal* rather than *natural* [1].

Most of the current efforts to cope with these problems are based on the use of previously developed *text-input* LT or LU systems relying on knowledge-based technology, which are serially coupled to the output of state-of-the-art word recognizer front-ends [2, 3, 4, 5]. Such a procedure is quite sensitive to front-end errors, since it does not exploit the powerful intrinsic restrictions that underly 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 LU 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 [1].

In this paper we describe how these goals can be achieved for *limited-domain* applications, and present comparative experimental results.

2. OVERVIEW OF THE RECOGNITION AND TRANSLATION SYSTEM

The purpose of a CSR system is to *translate* an acoustic signal into the word sequence uttered by the speaker. Current CSR systems use different knowledge sources modeling the successive mappings from the acoustic signal into phonemes, words and syntactically-correct sentences.

It is well known that when enough computational resources are available, the best strategy is to embed all these knowledge sources into an integrated model, and perform an optimal search for the best syntactically-

[1] Work partially supported by the Spanish CICYT under grant TIC92-1026-C02.

[2] Supported by the Spanish *Conselleria d'Educació i Ciència de la Generalitat Valenciana.*

[3] Supported by the Spanish *Ministerio de Educación y Ciencia.*

0-7803-2431-5/95 $4.00

correct word sequence given the acoustic signal. In this way, the restrictions imposed by the higher-level models allow for limiting the search space at the lower levels. This can be accomplished easily if (stochastic) finite-state models are chosen. The resulting integrated model is a graph through which an optimal path can be found by Dynamic Programming.

In Spoken LT (and, in particular, LU in the sense mentioned above), an additional mapping is required, from the word sequence uttered by the speaker into a sentence of a different language. If this mapping can be described by a finite-state model (which is often the case in limited-domain applications), the same techniques used to develop CSR systems can be applied. For this reason we chosed to work with *subsequential transducers* (SSTs), which have the additional advantage of being automatically learnable from training data [6].

A fully trainable Speech Translation and Understanding System has been developed [7]. This system is based on conventional *Viterbi beam search* through a network wich embeds phonetic, lexical and translation stochastic finite-state models. Phonetic models are discrete Hidden Markov Models. Lexical models describe words in terms of valid concatenations of phonemes. Translation models are SSTs, possibly embedding stochastic (finite-state) language models describing sentences in terms of possible concatenations of words.

In order to perform speech-input LT (and LU), two different approaches, hereafter referred to as *decoupled* and *integrated*, have been implemented and compared.

In the decoupled approach, a front-end speech recognizer is used, which is guided by phonetic, lexical and syntactic models of the input language. Translation is performed with text-to-text SSTs.

In the integrated approach, syntactic models of the input and/or output languages are used during the *training* of the SSTs, in order to obtain translation models compatible with the corresponding syntactic restrictions. During the recognition phase, these translation models are directly embedded with the phonetic and lexical models of the input language, in order to perform a global search for the optimal translation given the acoustic signal.

Among the different system components, probably the least known are the subsequential transducers. The next section is devoted to give an outline as well as appropriate references where more details can be found.

3. SUBSEQUENTIAL TRANSDUCER LEARNING

A *subsequential transducer* is a deterministic finite-state network that accepts sentences from a given input language and produces associated sentences of an output language. There is an input symbol and an output substring associated to each edge of a SST. Each state may also have associated an output substring. One of the states is the initial state and all the states are final. An input string is accepted if its sequence of symbols matches the associated input symbols of a sequence of edges. Simultaneously, an output string is produced which consists of the concatenation of the output substrings associated to the edges and to the last state used to accept s [8].

Given a set of training sentences from an unknown translation task, the *Onward Subsequential Transducer Inference Algorithm* (OSTIA) efficiently learns a SST that generalizes the training set [6]. Moreover, if the unknown target translation can be assumed to exhibit a subsequential structure, convergence to this translation is guaranteed if the set of training samples is "representative" or, simply, large enough [6].

This algorithm tends to generalize as much as possible while not contradicting the training data. While this has no negative effect if new correct (text) input sentences are submitted to translation, the results can be very bad if erroneous input data is used [7, 9]. 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-recognized sentences.

A recently introduced extended version of OSTIA [9] uses syntactic restrictions of the input and/or output languages, expressed by finite-state models, to constrain possible over-generalizations from the training data. It produces a subsequential transducer that only accepts input sentences and only produces 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 [9].

4. EXPERIMENTAL COMPARISON

4.1. Visual Scenes Description Task

The system has been tested with a pseudo-natural task recently proposed by Feldman et al [10]. This task consists of describing simple two-dimensional visual scenes which involve a few geometric objects with different

Spanish:	se añade un circulo grande y oscuro muy por encima del cuadrado pequeño y oscuro y del triangulo claro
English:	a large dark circle is added far above the small dark square and the light triangle
German:	man hat einen grossen dunklen Kreis weit uber dem kleinen dunklen Viereck und dem weissen Dreieck hinzugefügt
Semantic:	La(x) & D(x) & C(x) & Sm(z) & D(z) & S(z) & Li(w) & T(w) & FA(x;z) & FA(x;w) & Ad(x)

Figure 1: An example of translations of a Spanish sentence of the experimental task.

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 [11, 12]. 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. Examples of these input and output sentences are shown in Figure 1.

4.2. Training the acoustic, syntactic and translation models

For the experimental results reported below, phonetic models consist of 26 context-independent discrete Hidden Markov Models with 3 states and 128 codewords. They were trained using a small corpus of 120 sentences (from a different task) uttered by 10 speakers. Lexical models describing Spanish words (a total number of 29) consist of simple phoneme concatenations. The syntactic restrictions of the input and/or output languages have been modelized using stochastic *k-Testable Automata* (k-TA), which are equivalent to *k-Grams* [13, 14, 15, 16].

A set of 50100 input/output paired (text) sentences (for each of the 3 different output languages) was obtained using a semi-automatic procedure [11]. 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 subsequential transducers. For the decoupled approach SSTs where learnt with the original OSTIA (without input or output syntactic restrictions). For the integrated approach SSTs were learnt with the extended OSTIA using the input and/or output k-TA, 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 4 speakers (one of them also participated in the training of the HMMs). The system outlined above has been used to analyze these utterances, using the same beam search thresholds in all the experiments.

Figure 2 presents the recognition and translation word error rates (including insertion, substitution and deletion errors), averaged over the 4 speakers. An improvement of the results is observed as more syntactic constraints are integrated in the learnt transducers.

In the case of integrated transducers, only the output-language translation is directly available. However, a corresponding input-language sentence can be easyly obtained as a by-product; this enables to measure a "recognition error rate" in this case. It is worth noting that these *recognition* results (rows ii and iii) are significantly better than those achieved using the corresponding k-TA of the input language in the decoupled way (row i). This means that the (input parts of the) *transducers* learnt using these k-TAs offer better modeling of the input language than the k-TAs themselves.

Clearly, the use of more powerful acoustic models would improve both the recognition and the translation rates. However, results in row (i) reflect the fact that the original transducers (without models of the input or output languages) obtain relatively meaningless translations for incorrectly recognized sentences. Rows (ii) and (iii) show how this undesirable behaviour improves dramatically when integrated models are used.

The size (number of arcs) of the integrated SSTs was typically up to 5 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. The semantic representation was specifically chosen for studying this effect. For instance, in Figure 1 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.

Recog.	$k=2$	$k=3$	$k=4$
(i)	6.8 %	6.0 %	4.8 %
(ii)	4.8 %	4.2 %	3.5 %
(iii)	3.6 %	3.2 %	2.6 %

Trans.	$k=2$	$k=3$	$k=4$
(i)	27.4 %	23.2 %	15.2 %
(ii)	6.1 %	4.6 %	3.7 %
(iii)	3.9 %	3.3 %	2.8 %

(a) Spanish-English models

Recog.	$k=2$	$k=3$	$k=4$
(i)	6.8 %	6.0 %	4.8 %
(ii)	3.9 %	3.8 %	3.1 %
(iii)	3.3 %	2.6 %	2.3 %

Trans.	$k=2$	$k=3$	$k=4$
(i)	30.6 %	25.6 %	17.7 %
(ii)	5.6 %	5.8 %	4.7 %
(iii)	4.2 %	3.4 %	3.2 %

(b) Spanish-German models

Recog.	$k=2$	$k=3$	$k=4$
(i)	6.8 %	6.0 %	4.8 %
(ii)	4.4 %	3.8 %	3.4 %
(iii)	3.5 %	2.7 %	2.5 %

Trans.	$k=2$	$k=3$	$k=4$
(i)	31.2 %	25.9 %	16.4 %
(ii)	7.1 %	5.4 %	4.3 %
(iii)	4.6 %	3.5 %	3.9 %

(c) Spanish-Semantics models

Figure 2: Results with speech input: (i) Decoupled scheme: recognition guided by the k-TA of the input language, and translation performed with the transducers learnt by the original OSTIA; (ii) Integrated scheme: recognition and translation guided by the transducers learnt by the extended OSTIA using the k-TA of the input language only; (iii) Integrated scheme: recognition and translation guided by the transducers learnt by the extended OSTIA using both the k-TA of the input and output languages.

5. CONCLUDING REMARKS

Automatic translation (or understanding) of unrestricted spontaneous speech 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. For these kind of tasks, it is actually feasible to achieve a finite-state modeling not only of the syntactic constraints of the languages involved, but also of the required translation mapping itself. Moreover, since all the models can be automatically learnt from training data, for such a kind of tasks Speech Translation and Understanding Systems can now be built with a low development cost.

Clearly, the bottleneck of this approach lies in the availability of large corpora of paired sentences of different languages. It is likely that these resources will be available in the very near future, in the same way as corpora for training acoustic and language models exist today. An interesting feature of OSTIA is that the training sentences do not need to be aligned at any sub-sentence level. The algorithm implicitly learns the best possible alignments.

6. REFERENCES

[1] 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.

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

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

[4] 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.

[5] M. RAYNER, I. BRETAN, D. CARTER: "Spoken Language Translation with mid-90's Technology: A Case Study", Proc. of *EUROSPEECH-93*, pp. 1299–1302. Berlin (Germany), September 1993.

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

[7] 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*, H. Niemann, R. de Mori and G. Hanrieder (eds.), pp. 262–265. Infix, 1994.

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

[9] 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.

[10] 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.

[11] 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.

[12] A. CASTELLANOS, E. VIDAL, J. ONCINA: "Language Understanding and Subsequential Transducer Learning". Proc. of *1st International Colloquium on Grammatical Inference*, pp. 11/1–11/10, Colchester, England, 1993.

[13] 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.

[14] 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*, J. Rubio and J.M. López (eds.), Springer-Verlag, 1994.

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

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