The Data Driven Approach Applied to the OSTIA Algorithm*

José Oncina

Universidad de Alicante,
Departamento de Lenguajes y Sistemas informticos,
E-03071 Alicante (Spain)
oncina@dlsi.ua.es

Abstract. The OSTIA (Onward Subsequential Transducer Inference Algorithm) is an algorithm for inferring mappings between languages from input-output pairs, wich identifies in the limit any total subsequential function. It has been applied over a wide number of machine translation problems with great success. Incorporating the suggestions made in De la Higuera, Vidal and Oncina [dOV96] for automata inference, the DD-OSTIA (Data Driven OSTIA) is presented here. The experiments show a great reduction of the size of the training set needed for obtaining good models.

1 Introduction

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

Although natural languages are complex, the mappings defined by translations between them can be comparatively much simpler, specially when these languages are close each other as many European languages, and corpus-based (CB) techniques can be applied. There are some works that directly aim at placing Machine Translation (MT) within the CB framework ([BPPM93], [OGV91] and [VPL93]). Among these works, we will focus on one which is based on learning formal transducers. A formal transducer is a device that

* Work partially supported under grant TIC97-0941

accepts sentences from a given input and produces associated sentences of an output language. Formal transducers very seldom appear as components of MT systems. In fact, input-output relationships underlying natural language translation are very complex and the manual building of a transducer to account for these relationships, even in limited domain tasks, would be a too difficult or impossible task. However, some results in Transducer Learning have shown that this manual work can be avoided if a particular class of transducers is used, namely subsequential transducers. The *Onward Subsequential Transducer Inference Algorithm* (OSTIA) allows for the automatic learning of the structure of a (possibly very complex) transducer from a (perhaps large) set of training data consisting of input-output sentences [OGV91]. The performance of OSTIA can be significantly improved using the suggestions made in [dOV96] for learning Deterministic Finite Automata.

2 The algorithm

Formal descriptions of the OSTIA appeared elsewhere [Onc91] [OGV91] and here only an informal outline will be given.

A subsequential transducer is a finite-state network in which each edge has an *input symbol* and an *output string* associated to it. Each state may also have an output string associated to it. One of the states is the *initial state* and all the states can be *final* or *accepting*. For each state there is at most an outgoing edge for each symbol. An input string is *accepted* if its sequence of symbols matches the associated input symbols of a sequence of edges starting from the initial state. Every time an input string s is accepted, an output string is produced which consists on the concatenation of the output strings associated to the edges and the string associated to the last state used to accept s.

Two subsequential transducers are equivalent if they perform the same input-output mapping. For any subsequential transducer it is always possible to find an equivalent transducer that has the output strings assigned to the edges and states in such a way that they are as "close" to the initial state as they can be. This is called the *Onward Subsequential Transducer*.

In order to learn a subsequential transducer, OSTIA takes a finite training set of input-output pairs of sentences, T, as input, and proceeds in tree stages:

1. A prefix tree representation of all the input sentences of T is built. Then, null strings are assigned as output strings to both the internal nodes and the edges of this tree, while every output sentence of T is associated as a whole to the corresponding leaf of the tree. The result is called *Tree Subsequential Transducer* (TST).
2. By systematically moving the longest common prefixes of the output strings, level by level, from the leaves of the tree towards the root, an *Onward Tree Subsequential Transducer (OTST)*, equivalent to the TST is obtained.
3. Starting from the root, all pairs of states of the OTST are orderly considered, level by level, and they are (recursively) merged if this merging is *acceptable*; i.e., if the resulting transducer is subsequential and is not in contradiction with T. This can be checked by testing some conditions on the edges and states involved in the merge and their associated input symbols and output strings. Sometimes, "pushing back" some output strings toward the leaves of the tree is required in order to try to make a state merging acceptable.

All these operations can be very efficiently implemented, yielding a very fast algorithm that can easily handle huge sets of training data. At the end of the process, an Onward Subsequential Transducer which is a compatible generalization of T is obtained. It has been shown formaly that such strategy converges in the limit (that is, as the number of training pairs is sufficiently large) to the target subsequential transduction [Onc91].

Subsequential transducers and OSTIA have been successfully used so far in a variety of simple applications, some of which are quite contrived, such as learning to translate Roman numbers into their decimal representation, numbers written in English into their Spanish spelling, etc. [Onc91]. They have also been successfully applied to limited-domain language understanding, both in pseudo-natural and natural tasks like ATIS [CVO93], [PLV93]. Finally, some of these results have been extended by applying OSTIA to learn to translate (pseudo-natural) Spanish sentences describing simple visual scenes into corresponding English and German [CGV94].

The DD-OSTIA (Data Driven OSTIA) is inspired in the suggestion made by de la Higuera [dOV96] of substituting the order of considering the merge of states from the lexicogrphic order (by levels) to a heuristic order based on some measure of the equivalence of the states.

In the DD-OSTIA two mutual excluding subsets of states are defined in the OTST. In the beginning, the consolidated (C) subset only contains the initial state, and the frontier (F) subset contains all the states (not in C because the OTST is a tree) that are directly reachable (using an edge only) from a state in C.

Only pairs of states (c, f) such that $c \in C$ and $f \in F$ are considered. Given a state $f \in F$ such that there is no $c \in C$ such that c and f can be merged, then f is added to C and all the states directly reachable from f (using one edge only) are added to C. Otherwise, the pair of states (c, f) that maximizes the equivalence measure are merged.

This process is repeated until $F = \emptyset$. The equivalence measure that we have chosen is the reduction on the number of output symbols in the representation of the transducer if a tentative merge of both states is computed.

3 Experiments

The algorithm has been tested with an extension [CGV94] of a pseudo-natural task proposed by Feldman et al. [FLSW90]. 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 to Language Translation experimentation [CGV94], [CVO93]. Nowadays this corpus is becoming a benchmark for testing language translation techniques. Some examples of sentences of this corpus can be seen in fig. 1.

For the Spanish to English task Lozano [Loz96], using a grammar association [VPL93] with Markov Models technique reported a

Spanish:	`un cuadrado grande y un circulo oscuro estan muy a la derecha de un triangulo mediano y oscuro y un circulo`
English:	`a large square and a large circle are far to the right of a medium dark triangle and a circle`
Spanish:	`se elimina el circulo oscuro que esta debajo del circulo y del triangulo`
English:	`the dark circle which is bellow the circle and the triangle is removed`
Spanish:	`se añade un circulo pequeño y oscuro muy por encima del circulo mediano y oscuro y del circulo mediano`
English:	`a small dark circle is added far above the medium dark circle and the medium circle`

Fig. 1. Some pair of sentences for the Spanish to English translation from the Extended Feldman task.

22% of correct transduction using a training set of 1.000 pairs. This rate grows to 58% using 10.000 pairs. Using a grammar association technique with an association matrix the rates where 20% and 85% for 1.000 and 10.000 pairs respectively. Castaño and Casacuberta [CC97], using recurrent neural networks reported 53.1% and 98.4% for 500 and 3.000 pairs respectively. Prat [Pra98], using a grammar association technique [VPL93] reports, using 3.000 training pairs and using an improvement of the original technique a 81.6%, using the model 1 of IBM a 85.6%, using a multilayer perceptron a 95.% and using the Loco-C techniques a 95.8%.

In this section we perform these experiments for the Spanish to English translation. A series of increasing size random training sets, each including the previous one plus 100 new pairs, were used. Every set was submitted to the OSTIA and the DD-OSTIA and each subsequential transducer was used to translate 10.000 independent sentences. The results of the experiments appear in fig. 2.

It can be seen that using 3.000 training pairs the success ratio is 94.52%. Moreover, other sort of techniques (restrictions in the domain and range [OMV96]) can be applied in order to improve this rates.

4 Conclusions and future work

The use of a data driven strategy has shown to be fruitful when applied to the inference of subsequential transducers using OSTIA.

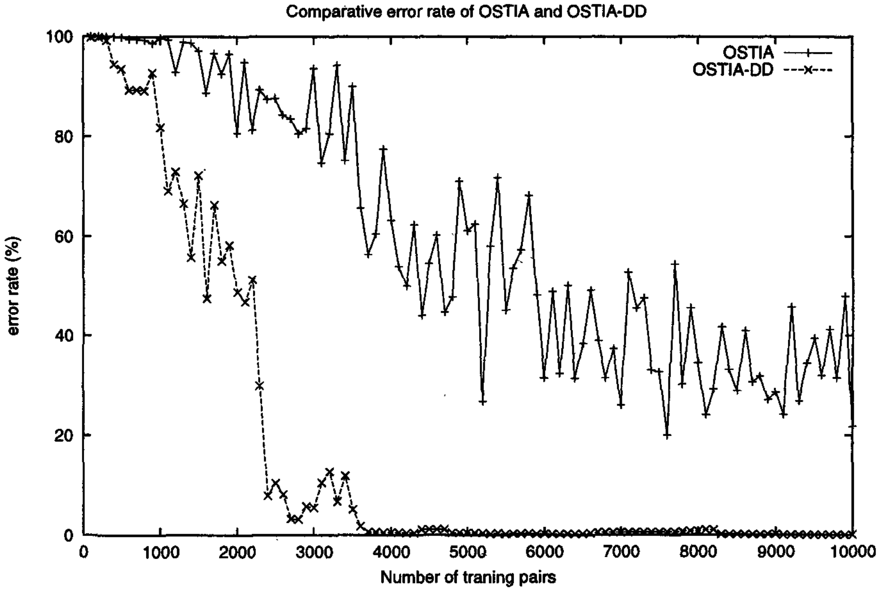

Fig. 2. Comparison of the error rates of the OSTIA and the DD-OSTIA when translating from Spanish to English within the extended Feldman task

One gets a dramatic reduction on the training set in order to obtain accurate models.

This technique can be merged with a technique of domain and range restriction [OMV96]. Other equivalence functions will be explored.

Acknowledgements: I thank M. Forcada and R.C.Carrasco for coments.

References

[BPPM93] 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*, 19:263–311, 1993.

[CC97] M. A. Castaño and F. Casacuberta. A connectionist approach to machine translation. In *Proceedings of the EuroSpeech'97*, volume 1, pages 91–94, Rodas, Greece, 1997. EuroSpeech'97.

[CGV94] A. Castellanos, I. Galiano, and E. Vidal. Applications of ostia to machine translation tasks. In *Grammatical Inference and Applications*, Lecture Notes in Artificial Intelligence, pages 93–105, Campello, Alicante, Spain, september 1994. 2nd International Colloquium on Grammatical Inference, Springer-Verlag.

[CVO93] A. Castellanos, E. Vidal, and J. Oncina. Language understanding and subsequential transducer learning. In *1nd International Colloquium on Grammatical Inference*, pages 11/1–11/10, Clochester, England, 1993.

[dOV96] C. de la Higuera, J. Oncina, and E. Vidal. Identification of dfa: Data-dependent versus data-independent algorithms. In *Grammatical Inference: Learning Syntax from Sentences*, Lecture Notes in Artificial Intelligence, pages 313–325, Montpellier, France, september 1996. 3rd International Colloquium on Grammatical Inference, Springer-Verlag.

[FLSW90] J. A. Feldman, G. Lakoff, A. Stolke, and S. H. Weber. Miniature language acquisition: A touchstone for cognitive science. Technical Report TR-90-009, Internationa Computer Science Institute, Berkeley, CA, USA, 1990.

[Loz96] M. Lozano. Asociación de gramaticas mediante modelos ocultos de markov. Master's thesis, Facultad de Informática, Universidad Politécnica de Valencia, Valencia, Spain, 1996.

[OGV91] J. Oncina, P. Gracía, and E. Vidal. Learning subsequential transducers for pattern recognition interpretation tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):252–264, March 1991.

[OMV96] J. Oncina and 96 M.A. Varó. Using domain information during the learning of a subsequential transducer. In *Grammatical Inference: Learning Syntax from Sentences*, Lecture Notes in Artificial Intelligence, pages 301–312, Montpellier, France, september 1996. 3rd International Colloquium on Grammatical Inference, Springer-Verlag.

[Onc91 J. Oncina. *Aprendizaje de lenguajes regulares y transducciones subsecuenciales.* PhD thesis, Universidad Politécnica de Valencia, Valencia, Spain, 1991.

[PLV93] R. Pieraccini, E. Levin, and E. Vidal. Learning how to understand language. In *Proceedings of the EuroSpeech'93*, volume 2, pages 448–458, Berlin, Germany, September 1993. 3rd European Conference on Speech Communication and Technology.

[Pra98] F. Prat. *Traducción automática en dominios restringidos: Algunos modelos estocásticos susceptibles de ser aprendidos a partir de ejemplos.* PhD thesis, Universidad Politécnica cd Valencia, Valencia, Spain, 1998.

[VPL93] E. Vidal, R. Pieraccini, and E. Levin. Learning associations between grammars: A new approach to natural language understanding. In *Proceedings of the EuroSpeech'93*, volume 2, pages 1187–1190, Berlin, Germany, September 1993. 3rd European Conference on Speech Communication and Technology.