Mach Learn (2007) 66:69–91
DOI 10.1007/s10994-006-9612-9

Learning finite-state models for machine translation

Francisco Casacuberta · Enrique Vidal

Received: 30 June 2005 / Revised: 12 April 2006 / Accepted: 22 June 2006 / Published online: 4 August 2006

Abstract In formal language theory, finite-state transducers are well-know models for simple "input-output" mappings between two languages. Even if more powerful, recursive models can be used to account for more complex mappings, it has been argued that the input-output relations underlying most usual natural language pairs can essentially be modeled by finite-state devices. Moreover, the relative simplicity of these mappings has recently led to the development of techniques for learning finite-state transducers from a training set of input-output sentence pairs of the languages considered. In the last years, these techniques have lead to the development of a number of machine translation systems. Under the statistical statement of machine translation, we overview here how modeling, learning and search problems can be solved by using stochastic finite-state transducers. We also review the results achieved by the systems we have developed under this paradigm. As a main conclusion of this review we argue that, as task complexity and training data scarcity increase, those systems which rely more on statistical techniques tend produce the best results.

Keywords Machine translation · Stochastic finite-state transducers · Grammatical inference

1. Introduction

Machine translation (MT) is one of the most appealing (and challenging) applications of human language processing technology. Owing to its great social and economical interest, in the last 20 years MT has been considered under almost every imaginable point of view: from strictly linguistics-based methods to pure statistical approaches including, of course,

This work was partially supported by the European Union project TT2 (IST-2001-32091) and by the Spanish project ITEFTE (TIC 2003-08681-C02-02).

Editor: Georgios Paliouras and Yasubumi Sakakibara

F. Casacuberta (✉)· E. Vidal
Departamento de Sistemas Informáticos y Computación, Instituto Tecnológico de Informática,
Universidad Politécnica de Valencia, 46071 Valencia, Spain
e-mail: fcn@dsic.upv.es

formal language theory and the corresponding learning paradigm, *grammatical inference* (GI). Different degrees of success have been achieved so far using these approaches.

Basic MT consists in transforming text from a source language into a target language, but several extensions to this framework have been considered. Among the most interesting of these extensions are *speech-to-speech* MT (STSMT) and *computer assisted* (human) *translation* (CAT). In STSMT, which is generally considered significantly harder than pure text MT, the system has to accept a source-language utterance and produce corresponding human-understandable target-language speech. In CAT, on the other hand, the input is source-language text and both the system and the human translator have to collaborate with each other in an attempt to produce high quality target text.

Nowadays, the so called *statistical approach to machine translation* (SMT) is one of the most successful approaches. This success is mainly due to the fact that the corresponding models can be automatically learned from bilingual training data (Brown et al., 1990, 1993). Basically, the models in SMT aim to characterize the statistical relations between pairs of single words (Brown et al., 1993; Ney et al., 2000) or sequences of words (or *phrases*) (Tomás & Casacuberta, 2001; Och & Ney, 2004; Zens & Ney, 2004) from the languages considered. One of the main problems with such models is the lack of efficient algorithms both for learning and for performing the translation of new sentences (searching). In practice, for the most sophisticated models, only heuristic approximations have been implemented for training (Brown et al., 1993; Ney et al., 2000), based on (approximations to) the *expectation–maximization* (EM) algorithm (Och & Ney, 2003). The search with these models is also a computationally difficult problem (Knight, 1999) and only suboptimal solutions are available. Some of these approximations are based on A*-like techniques (Berger et al., 1996), greedy algorithms (Germann, 2003) and/or dynamic-programming strategies (Tillmann & Ney, 2003). Syntactic aspects of the language can also be incorporated in this framework (Charniak, Knight, & Yamada, 2003).

Several public-domain toolkits are available for training these statistical models: the first one was EGYPT, developed by the statistical machine translation team during the summer workshop in 1999 at the Center for Language and Speech Processing, Johns-Hopkins University (Knight et al., 1999). An improved version of this toolkit is GIZA++ (Och & Ney, 2003). Recently, the toolkit THOT has been presented for learning phrase-based models (Ortiz, García-Varea, & Casacuberta, 2005). On the other hand, toolkits implementing search engines for translation are scarce; an example is PHARAOH, a beam search engine for phrase-based models (Koehn, 2004).

We consider here MT, STSMT and CAT models that can be be automatically learned through suitable combinations of GI and statistical methods from training pairs of sentences. More specifically, we focus on *stochastic finite-state transducers* (SFST), which constitute an important family of models within the theory of formal languages (Vidal et al., 2005). With respect to SMT, SFSTs offer the availability of more efficient searching algorithms and the promise of effective, possibly faster learning techniques. On the other hand, SFSTs permit a simple integration with other information sources (such as acoustic models), which makes it easy to apply SFSTs to more difficult tasks such as speech translation (Casacuberta et al., 2004). SFSTs and the corresponding training and search techniques have been studied by several authors, in many cases explicitly motivated by MT applications (Vidal, García, & Segarra, 1989; Oncina, García, & Vidal, 1993; Knight & Al-Onaizan, 1998; Alshawi, Bangalore, & Douglas, 2000; Bangalore & Riccardi, 2003; Kumar & Byrne, 2003; Casacuberta & Vidal, 2004; Tsukada & Nagata, 2004; Kumar, Deng, & Byrne, 2006).

Other models that can be considered generalizations of STSTs are the *inversion transduction grammars* (Wu, 1995) and the *head transducers* (Alshawi, Bangalore, & Douglas, 2000).

These models are theoretically more powerful than SFSTs, but they are more expensive from a computational point of view.

This paper summarizes our work for more than ten years of applied research on SFST learning, mainly aimed at MT applications, which has been supported by several projects funded by National and European agencies, as well as by private companies (Tracom, 1995; Extra, 1997; EuTrans, 2000; SisHiTra, 2001; TransType2, 2001; TeFaTe, 2003). The developments carried out in these projects have always been competitive with other state-of-the-art technologies: from *knowledge-based* and *memory-based*, to *"pure" statistical* techniques. The MT, STSMT and CAT tasks considered in these works have entailed ever increasing demands in task complexity, i.e. larger vocabularies, less restricted domains, more difficult pairs of languages and relatively smaller amounts of training data. So, the present article is devoted to review how our initial, GI-based, SFST learning technology has evolved in order to cope with these growing challenges.

The general (probabilistic) statement of MT problems is presented in Section 2 and the SFSTs are introduced in Section 3. A simple taxonomy of the techniques to infer SFST is introduced in Section 4. Some of these techniques are reviewed in Sections 5 and 6. Section 7 briefly overviews one of the most popular "pure" statistical techniques used in MT, for which some comparative results are presented in Section 9. A summary of the main experiments carried out and the corresponding results are presented in Sections 8 and 9, respectively. Finally, some conclusions are drawn in Section 10.

2. General statement of MT problems

The (*text-to-text*) MT problem can be statistically stated as follows. Given a sentence $\mathbf{s}$ from a source language, search for a target-language sentence $\hat{\mathbf{t}}$ which maximizes the posterior probability:[1]

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \Pr(\mathbf{t}|\mathbf{s}). \tag{1}$$

It is commonly accepted that a convenient way to deal with this equation is to transform it by using the Bayes' theorem:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \Pr(\mathbf{t}) \cdot \Pr(\mathbf{s}|\mathbf{t}), \tag{2}$$

where $\Pr(\mathbf{t})$ is a target *language model* –which gives high probability to well formed target sentences– and $\Pr(\mathbf{s}|\mathbf{t})$ accounts for source-target word(-position) relations and is based on *stochastic dictionaries* and *alignment models* (Brown et al., 1993; Ney et al., 2000; Och & Ney, 2003). These models have been extended to model alignments between word sequences instead of single words (Tomás & Casacuberta, 2001; Marcu & Wong, 2002; Zens, Och, & Ney, 2002; Och & Ney, 2004).

Alternatively, the conditional distribution in Eq. (1) can be transformed into a joint distribution:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \Pr(\mathbf{s}, \mathbf{t}), \tag{3}$$

[1] For simplicity, $\Pr(X = x)$ and $\Pr(X = x \mid Y = y)$ are denoted as $\Pr(x)$ and $\Pr(x \mid y)$.

which, among other possibilities, can be adequately modeled by means of SFSTs. This is the kind of models considered in the present work. As previously mentioned, one attractive feature of these models is that a good approximate solution to Eq. (3) can be efficiently obtained by the well-known *Viterbi algorithm* (Picó & Casacuberta, 2001).

Let us now consider the STSMT problem. Here, an acoustic representation of a source-language utterance $\mathbf{x}$ is available and the problem is to search for a target-language sentence $\hat{\mathbf{t}}$ that maximizes the posterior probability:[2]

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \operatorname{Pr}(\mathbf{t} \mid \mathbf{x}). \tag{4}$$

Every possible decoding of a source utterance $\mathbf{x}$ in the source language can be considered as the value of a hidden variable, $\mathbf{s}$ (Casacuberta et al., 2004). Therefore, assuming that $\operatorname{Pr}(\mathbf{x}|\mathbf{s}, \mathbf{t})$ does not depend on $\mathbf{t}$, Eq. (4) can be rewritten as:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \sum_{\mathbf{s}} \operatorname{Pr}(\mathbf{s}, \mathbf{t}) \cdot \operatorname{Pr}(\mathbf{x}|\mathbf{s}). \tag{5}$$

As in plain MT, $\operatorname{Pr}(\mathbf{s}, \mathbf{t})$ can be modeled by a SFST. The term $\operatorname{Pr}(\mathbf{x}|\mathbf{s})$, on the other hand, can be modeled through *hidden Markov models* (HMM), which are the standard acoustic models in automatic speech recognition (Jelinek, 1998).

Most STSMT systems presented in the last few years are based on a trivial architecture, often called *"serial"* or *"loosely-coupled"*: Source-language speech is first decoded into a sequence of source-language words, using a standard automatic speech recognizer and the resulting source word sequence is translated into a target sentence as in conventional text-to-text MT. This architecture constitutes a very crude approximation to Eq. (5) (Casacuberta et al., 2004). However, if the underlying source–target mapping is modeled by a SFST, the *integrated architecture*, where all the acoustic, syntactic and translation models are tightly combined into a single (maybe huge) finite-state model, allows for a potentially better approximation. Moreover, the *Viterbi algorithm* can also be used thanks to the homogeneous finite-state nature of both SFST and HMMs.

Finally, let us consider a simple statement of CAT (Langlais, Foster, & Lapalme, 2000). Given a source text $\mathbf{s}$ and a fixed *prefix* of the target sentence $\mathbf{t}_p$ –previously validated by the human translator–, the problem is to search for a *suffix* of the target sentence $\hat{\mathbf{t}}_s$ that maximizes the posterior probability:

$$\hat{\mathbf{t}}_s = \underset{\mathbf{t}_s}{\operatorname{argmax}} \operatorname{Pr}(\mathbf{t}_s | \mathbf{s}, \mathbf{t}_p). \tag{6}$$

This equation can be equivalently written as:

$$\hat{\mathbf{t}}_s = \underset{\mathbf{t}_s}{\operatorname{argmax}} \operatorname{Pr}(\mathbf{s}, \mathbf{t}_p, \mathbf{t}_s). \tag{7}$$

Equation (7) is similar to Eq. (3), but here the maximization is constrained to a set of suffixes, rather than full sentences. As in Eq. (3), this joint distribution can be adequately modeled by means of SFSTs and the Viterbi algorithm can be adequately adapted for searching (Civera et al., 2004).

[2] From $\hat{\mathbf{t}}$, a *text-to-speech* synthesizer can be used to produce a target utterance.

All the above problem statements share the common *learning problem* of estimating $\Pr(\mathbf{s}, \mathbf{t})$, which can be approached by training a SFST from a parallel text corpus. This will be discussed in detail in the following sections.

3. Stochastic finite-state transducers

Stochastic finite-state transducers (SFST) are similar to stochastic finite-state grammars or automata, but in this case two different alphabets are involved: Source and target alphabets. Each transition in a SFST has attached a source word and a (possible empty) string of target words.[3] A SFST $\mathcal{T}_P$ is defined as tuple $\langle \Sigma, \Delta, Q, q_0, p, f \rangle$, where (Vidal et al., 2005):

- Σ is a finite set of *source (input) words*;
- Δ is a finite set of *target (output) words*;
- Q is a finite set of *states*
- $q_0 \in Q$ is the *initial state* and
- $p : Q \times \Sigma \times \Delta^\star \times Q \rightarrow [0, 1]$ is a *transition probability function*
- $f : Q \rightarrow [0, 1]$ is a *final-state probability function*

The functions p and f must verify:

$$\forall q \in Q, \quad f(q) \quad + \sum_{(\mathrm{s}, \tilde{\mathrm{t}}, q') \in \Sigma \times \Delta^\star \times Q} p(q, \mathrm{s}, \tilde{\mathrm{t}}, q') = 1. \tag{8}$$

The non-probabilistic counterpart $\mathcal{T}$ of a given a SFST $\mathcal{T}_P$, called *characteristic finite-state transducer of* $\mathcal{T}_P$ (FST), can be defined. The *transitions* are those tuples in $Q \times \Sigma \times \Delta^\star \times Q$ with probability greater than zero and the *set of final states* are those states in Q with final-state probability greater than zero.

Given $\mathcal{T}_P$, a *translation form* with I transitions associated with the *translation pair* $(\mathbf{s}, \mathbf{t}) \in \Sigma^* \times \Delta^*$ is a sequence of transitions $\phi = (q_0, \mathrm{s}_1, \tilde{\mathrm{t}}_1, q_1)\ (q_1, \mathrm{s}_2, \tilde{\mathrm{t}}_2, q_2)\ (q_2, \mathrm{s}_3, \tilde{\mathrm{t}}_3, q_3) \ldots (q_{I-1}, \mathrm{s}_I, \tilde{\mathrm{t}}_I, q_I)$, such that $\mathrm{s}_1\, \mathrm{s}_2 \ldots \mathrm{s}_I = \mathbf{s}$ and $\tilde{\mathrm{t}}_1\, \tilde{\mathrm{t}}_2 \ldots \tilde{\mathrm{t}}_I = \mathbf{t}$. Its probability is the product of the corresponding transition probabilities, times the final-state probability of the last state in the sequence:

$$\Pr_{\mathcal{T}_P}(\phi) = \prod_{i=1}^{I} p(q_{i-1}, \mathrm{s}_i, \tilde{\mathrm{t}}_i, q_i) \cdot f(q_I). \tag{9}$$

The set of translations forms associated with a translation pair $(\mathbf{s}, \mathbf{t})$ with probability higher than zero is denoted as $\Phi(\mathbf{s}, \mathbf{t})$.

The *probability of a translation pair* $(\mathbf{s}, \mathbf{t})$ according to $\mathcal{T}_P$ is then defined as the sum of the probabilities of all the translation forms associated with $(\mathbf{s}, \mathbf{t})$:

$$\Pr_{\mathcal{T}_P}(\mathbf{s}, \mathbf{t}) = \sum_{\forall \phi \in \Phi(\mathbf{s}, \mathbf{t})} \Pr_{\mathcal{T}_P}(\phi)\,. \tag{10}$$

If $\mathcal{T}_P$ has no useless states, $\Pr_{\mathcal{T}_P}(\mathbf{s}, \mathbf{t})$ describes a probability distribution on $\Sigma^\star \times \Delta^\star$ which is called *stochastic finite-state translation*. This distribution is used to model the

[3] The term "word" is used to refer a single token as in MT i.e. a "symbol" in formal language theory.

joint probability introduced in Eq. (3) of the previous section. The terms *regular* or, more properly, *rational* translations are also often used in the scientific literature to refer to (the non-probabilistic counterpart of) these mappings (Berstel, 1979).

Any SFST has two embedded stochastic regular languages, one for the source alphabet and another for the target alphabet. These languages correspond to the two marginals ($\Pr_i$ and $\Pr_o$) of the joint distribution modeled by the the SFST:

$$\Pr_i(\mathbf{s}) = \sum_{\mathbf{t}\in\Delta^\star} \Pr(\mathbf{s},\mathbf{t}), \quad \Pr_o(\mathbf{t}) = \sum_{\mathbf{s}\in\Sigma^\star} \Pr(\mathbf{s},\mathbf{t}).$$

In practice, the corresponding source or target finite-state grammars are obtained from the finite-state transducer by dropping the target or source words of each transition, respectively.

A particularly interesting transducer model is the *subsequential transducer (SST)* which is a finite-state transducer with the basic restriction of being deterministic (Vilar, 2000). This implies that if two transitions have the same starting state and the same source word, then both ending states are the same state and both the target strings are also the same target string. In addition, SSTs can produce a target substring when the end of the input string has been detected.

SFSTs exhibit properties and problems similar to those exhibited by stochastic regular languages. One of these properties is the formal basis of the GIATI technique for transducer inference presented in Section 6.2. It can be stated as the following theorem (Casacuberta, Vidal, & Picó, 2005): *Every stochastic finite-state translation can be obtained from a stochastic regular language and two morphisms*. This is a weaker version of the stochastic extension (Casacuberta, Vidal, & Picó, 2005) of a classical morphism theorem (Berstel, 1979): *Every rational translation can be obtained from a local language and two alphabetic morphisms*, where a *local language* is defined by a set of permitted two-word segments (and therefore a *stochastic local language* is equivalent to a *bigram distribution* (Vidal et al., 2005)). In both cases, the morphisms allow for building the components of a pair of the finite-state translation from a string of the corresponding local language (Casacuberta, Vidal, & Picó, 2005).

By using a SFST $\mathcal{T}_P$ to approach the joint distribution $\Pr(\mathbf{s},\mathbf{t})$, Eq. (3) is rewritten as:

$$\hat{\mathbf{t}} = \operatorname*{argmax}_{\mathbf{t}\in\Delta^\star} \Pr_{\mathcal{T}_P}(\mathbf{s},\mathbf{t}). \tag{11}$$

That is, given $\mathcal{T}_P$ and $\mathbf{s}\in\Sigma^\star$, search for a target string $\hat{\mathbf{t}}$ which maximizes $\Pr_{\mathcal{T}_P}(\mathbf{s},\mathbf{t})$. While this *SFST stochastic translation problem* is proved to be **NP**-Hard (Casacuberta and de la Higuera, 2000), a generally good approximation can be obtained in polynomial time through a simple extension of the Viterbi algorithm (Picó & Casacuberta, 2001). This approximation consists in replacing the sum operator in Eq. (10) with the maximum operator:

$$\Pr_{\mathcal{T}_P}(\mathbf{s},\mathbf{t}) \approx \widehat{\Pr}_{\mathcal{T}_P}(\mathbf{s},\mathbf{t}) = \max_{\forall\phi\in\Phi(\mathbf{s},\mathbf{t})} \Pr_{\mathcal{T}_P}(\phi). \tag{12}$$

This approximation to the SFST stochastic translation problem permits the computation of the optimal translation form (with respect to Eq. (12)) in linear time with the number of source words. The translation of the given source sentence is then approached as the sequence of target strings which appear in this optimal translation form.

4. Learning stochastic finite-state transducers

Following the statistical framework adopted in the previous sections, three main families of techniques can be used to learn a SFST from a parallel corpus of source-target sentences:

- *Traditional syntactic pattern recognition paradigm*: (a) Learn the SFST structure or "topology" (the *characteristic transducer*) and (b) Estimate its probabilities from the same data.
- *Hybrid methods*: Under the *traditional* paradigm, use statistical methods to guide the structure learning.
- *Pure statistical approach*: (a) Adequately parameterize the SFST structure and consider it as a hidden variable and (b) estimate everything by expectation maximization (EM).

To estimate the probabilities in the *traditional* approach, *maximum likelihood* or other possible criteria can be used (Picó & Casacuberta, 2001). As in every estimation problem, an important issue is the modeling of unseen events. In our case, a general approach to this *smoothing* problem consists in using *stochastic error-correcting parsing* (Amengual & Vidal, 1998; Amengual et al., 2001). In this case, an explicit error model (for example, the classical substitution/insertion/deletion errors) is assumed and its parameters estimated from training data. Alternatively, smoothing techniques borrowed from the field of language modeling (Ney et al., 1997) can be used. There are two known possibilities to apply these techniques: Either within of the process of learning both the structural and probabilistic SFST components (Casacuberta & Vidal, 2004), or in the estimation of the probabilities of the SFST themselves (Llorens, Vilar, & Casacuberta, 2002). In the first case, the smoothed model generated within the learning process is exported to the final SFST. The second case is a generalization of the methods proposed for n-grams models. In each state of a n-gram model there is only one equivalent class of paths that arrive to the state. In the states of automata and finite-state transducers, there can be more than one equivalent class of paths.

We should mention that the computational demands of all the (structural) learning and (statistical) estimation techniques here considered is *low*; typically they exhibit (close to) *linear complexity* with the size of the training data. Given the huge training corpora generally needed in real MT applications, only low-cost learning and estimation algorithms can be afforded in practice. Therefore we will not further discuss these computational details. Instead, we will be interested in the relative amounts of training data required by the different techniques in order to achieve reasonable generalizations.

5. Traditional syntactic pattern recognition paradigm: OSTIA

The formal model of translation used in this section is the SST. Thanks to the capability of producing output substrings in its (final) states, a sub-sequential transducer can delay the output of target words until enough source words have been seen to determine a correct output. Therefore, the states of a SST hold the "memory" of the part of the sentence seen so far. This allows the whole context of a word to be taken into account, if necessary, for the translation of the next word. A very efficient technique for automatically learning these models from a training set of sentence pairs is the so called *onward subsequential transducer inference algorithm* (OSTIA) (Oncina, García, & Vidal, 1993).

OSTIA starts building up an initial representation of the original paired data in the form of a tree (the *onward prefix tree transducer*). Then appropriate states of this transducer are merged to build a FST in which those transitions sharing some structural properties are merged together in order to generalize the seen samples. To this end, the tree is traversed level by level and the states in each level are considered to be merged with those previously visited. Only those pairs of states which are compatible according to the output strings of their subtrees are effectively merged. If the training pairs were produced by an unknown SST, $\mathcal{T}$, –which can be considered true, at least for many common pairs of natural languages– and the amount of training data is sufficiently large and/or representative, then OSTIA is guaranteed to converge to a canonical (*onward*) SST which generates the same translation pairs as $\mathcal{T}$ (Oncina, García, & Vidal, 1993).

The state merging process followed by OSTIA tries to generalize the training pairs as much as possible. This often leads to very compact transducers which adequately translate correct source text of the learned task into the target language. However, this compactness often entails an excessive generalization of the source and the target languages, allowing meaningless source sentences to be accepted, and even worse target sentences to be produced. This is not a problem for perfectly correct source text, but becomes important when not exactly correct text or speech is to be used as input.

A possible way to overcome this problem consists in further restricting state merging so that the resulting SST only accepts source sentences and produces target sentences that are consistent with given source and target (regular) language models. These models are known as *domain* and *range* language models, respectively. A version of OSTIA called OSTIA-DR (Oncina & Varó, 1996; Castellanos et al., 1998) enforces these restrictions in the learning process. N-grams (Jelinek, 1998), generally trained from the source and target sentences in the given corpus, are usually adopted as domain and range language models for OSTIA-DR.

OSTIA and OSTIA-DR have been applied to many relatively simple MT tasks, including speech-input MT. Results will be reported in Section 9.

6. Hybrid methods: OMEGA and GIATI

OSTIA has proved able to learn adequate transducers for real (albeit limited) tasks if a sufficiently large amount of training pairs is available. However, as the amount of training data shrinks, its performance drops dramatically. Clearly, in order to convey enough information to learn structurally rich transducers, prohibitively large amounts of examples are required. Therefore, in order to render training data demands realistic, additional, explicit information about the relation of source-target words involved in the translation seems to be needed. The two techniques discussed in this section are explicitly based on this idea.

6.1. OMEGA

This algorithm, called “OSTIA modified for employing guarantees and alignments” (OMEGA) (Vilar, 2000), is an improvement over OSTIA-DR to learn SSTs.

As with OSTIA, there are two main training phases: building an initial tree from the training pairs and state-merging. Apart from the OSTIA-DR state-merging restrictions, including those derived from (n-gram) source and target language models, two additional knowledge sources can be employed to avoid overgeneralisation: a *bilingual dictionary* and/or *word alignments*. They are used to label each state of the initial tree with two sets of target words called *“guarantees”* and *“needs”*. The state-merging process is then further constrained by the information contained in these sets (Vilar, 2000). This ensures that target words will never be produced before having seen the source words they are translation of.

These dictionaries and alignments can be obtained from the training pairs by means of pure statistical methods such as those described in Brown et al. (1993) and Och and Ney (2003).

Results obtained using OMEGA both for MT and STSMT will be reported in Section 9. Overall, OMEGA has proved adequate to cope with low and moderate-difficulty tasks. However, as task complexity increases, results degrade rather sharply. This prompted the need for even less data-hungry transducer learning techniques, leading to the approach discussed below.

6.2. GIATI

The finite-state models underlying both OSTIA(-DR) and OMEGA rely on delaying the output of target words until enough source text have been seen to guarantee a correct output. In fact, every finite-state device must rely on a similar principle as applied to MT. This delaying process can be seen from a different point of view involving a *left-to-right bilingual segmentation* of the source and target sentences: Each sequence of source words we have to wait for, determines a *source segment*, while the corresponding sequence of target words constitutes the associated *target segment*.

For most language pairs of interest, these segments are typically rather short; less then five words on the average in many cases. Therefore, we can capitalize on this fact to develop transducer learning techniques that explicitly rely on left-to-right segmentations of the training pairs. While this idea can be considered unacceptable from a linguistic point of view, it is quite reasonable in practice. MT techniques can take advantage of the word-order flexibility allowed in many languages and of the fact that the system output sentences are aimed at being acceptable by human beings. For instance, an English sentence such as *"can you give me the keys to my room, please?"* can be translated into Spanish as *"por favor, ¿puede darme las llaves de mi habitación?"*, which entails a long-term reordering for the words *"por favor"* and *"please"*. This would prevent any left-to-right segmentation into bilingual segments shorter than the whole sentence pair. However, other adequate Spanish translations are possible, such as *"¿puede darme, por favor, las llaves de mi habitación?"*, or even *"¿puede darme las llaves de mi habitación, por favor?"*, which admit natural left-to-right segmentations into very short bilingual segments. Of course, this kind of flexibility does appear in most training sets, which adds to the difficulty of obtaining good generalizations from the given training pairs. However, as the amount of training data becomes sufficiently large, learning algorithms can find enough instances of the kind of data they can take advantage of in order to produce adequate generalizations.

One of the techniques explicitly based on these ideas is called *grammatical inference and alignments for transducer inference* (GIATI). It was first introduced in Casacuberta (2000) and further developed in Casacuberta, Vidal, and Picó (2005). As in the case of OMEGA, GIATI also makes use of information obtained by means of pure statistical methods. However, in this case, the alignments are used to define the required bilingual segmentations.

More formally speaking, given a finite sample of string pairs, GIATI relies on the two-morphisms theorem mentioned in Section 3, to propose the following steps for the inference of a finite-state transducer:

1. Using a left-to-right bilingual segmentation of each training pair of sentences, the pair is appropriately transformed into a single string from an *extended alphabet*, which is composed of pairs of source-target word sequences.
2. A stochastic regular grammar, typically an n-gram, is inferred from the set of strings obtained in first step.

3. The terminal words of the grammar rules from the second step are transformed back into source/target word strings by applying adequate morphisms. This converts the stochastic grammars into the learned transducer.

The main problem with this approach is the first (and correspondingly the third) step(s), i.e. to adequately transform a parallel corpus into a string corpus. The transformation of the training pairs must capture the correspondences between words of the input and the output sentences and must permit the implementation of the inverse transformation of the third step. As previously mentioned, this is achieved with the help of bilingual segmentation (Casacuberta & Vidal, 2004).

To illustrate this first step, we will use the pair (*"un coche rojo" , "a red car"*). A suitable word-alignment would align "*un*" with "*a*", "*coche*" with "*car*" and "*rojo*" with "*red*". From this alignment we could naively build the string "*(un,a) (coche,car) (rojo,red)*". However, this would imply a reordering of the words "*red*" and "*car*", difficult to model in the finite-state framework. Instead of reordering, the alignment can be used to produce a left-to-right bilingual segmentation into two segments: ("*un*" , "*a*") and ("*coche rojo*" , "*red car*"). This segmentation directly yields the single string and the corresponding extended alphabet required for GIATI. Nevertheless, by allowing empty target segments, a simpler segmentation which has proved equivalently adequate in practice, can be produced as follows: If the source segment of a pair has more than one word, the segment is further splitted into several single-word segments, each paired with an empty target string, except the last one, which is paired with the original whole target segment. In our example this results in the three bilingual segment string: "*(un,a) (coche,-) (rojo,red car)*".

The first step of the GIATI technique is based on *statistical alignment models* (Och & Ney, 2003) which are based on functions from source sentence positions to target sentence positions; that is, these models are based on *asymmetric* alignments. Some techniques to symmetrize the alignments have been proposed in Och and Ney (2004) and Tomás, Lloret, & Casacuberta (2005), but only marginal improvements were achieved by applying these techniques to GIATI (Picó, Tomás, & Casacuberta, 2004).

The probabilities associated with the transitions of a SFST learned in this way are just those of the corresponding stochastic regular grammar inferred in *step 2*. Therefore, an interesting feature of GIATI is that it can readily make use of all the smoothing techniques known for *n*-grams (Ney et al., 1997) and for stochastic regular grammars (Llorens, Vilar & Casacuberta, 2002). Note, however, that GIATI extended alphabets can be very large, specially for large vocabularies and/or language pairs where left-right alignments are unnatural. In some cases, this may entail severe a data sparseness, which makes the estimation problem difficult to solve, even with the help of the mentioned smoothing techniques. Other interesting properties of GIATI can be found in Casacuberta, Vidal, and Picó (2005).

GIATI has been applied to a variety of translation tasks, leading to good results in many cases. These results will be reported in Section 9.

7. Statistical alignment templates

In the experiments to be presented in the following sections, the results of the SFST approaches described so far will be compared with those obtained by one popular technique which is commonly considered to be among the best techniques available in the "pure" statistical framework. Here we will briefly outline this approach, which is called *statistical alignment templates* (Och & Ney, 2004). It can be considered as representative of the current tendencies of state-of-the-art statistical MT, where best results are typically obtained using techniques based in *phrases* rather than in individual words (Och & Ney, 2004; Tomás & Casacuberta, 2006).

In this approach, an entire group of adjacent words in the source sentence may be aligned with an entire group of adjacent target words. As a result, the word context has a greater influence and the changes in word order from source to target language can be learned explicitly. The alignment of word groups is carried out through *templates*. A template establishes the alignment (possibly by means of reordering) between two sequences of word classes, which are learned automatically using a bilingual corpus. However, the lexical model, which is an integrated part of each template, is still to be based on word-to-word correspondences (Och & Ney, 2004).

Learning such models consists of three phases: 1) Computation of word alignments (as in the GIATI approach); 2) extraction of bilingual phrases including phrase alignments and 3) generalization of these bilingual phrases to templates by including the word alignments within the templates and by using word classes. These models are usually combined with other complementary models (target language models, etc.) in a log-linear framework. The translation is performed by using a breadth-first graph search algorithm with pruning (Och & Ney, 2004).

8. Experimental framework

The techniques described in the previous sections have been applied to a large variety of MT, STSMT and CAT tasks. In this section the characteristics of the corpora used in the empirical tests with these applications are described, along with the assessment measures adopted to evaluate the quality of the inferred models.

8.1. Translation tasks and the corresponding corpora

Here we briefly review some of the MT tasks we have considered and the corresponding bi-text corpora. Table 1 summarizes the main features of these corpora and Table 2 shows corresponding examples of sentence pairs. Most of these corpora were produced or compiled to test the capabilities of systems developed in several MT projects. The data sets belong to

Table 1 Corpora used in the experiments. Languages: En = English, Sp = Spanish, It = Italian, Ba = Basque, Fr = French and Ge = German. Vocabulary sizes are the numbers of *different words*, while training and test sizes are the numbers of *running words*

Task	Languages	Vocabularies	Training sizes	Test sizes
MLT	Sp / En	29 / 25	128 K / 119 K	160 K /148 K
EUTRANS-0	Sp / En	0.7 K / 0.5 K	4.7 M / 4.8 M	35 K / 36 K
EUTRANS-I	Sp / En	0.7 K / 0.5 K	97 K / 99 K	35 K / 36 K
EUTRANS-II	It / En	2.5 K / 1.7 K	55 K / 64 K	6.1 K / 7.2 K
AMETRA-MET	Sp / Ba	0.7 K / 1.3 K	87 K / 77 K	1.3 K / 1.2 K
TT2-Xrce	En / Sp	8 K / 12 K	0.6 M / 0.7 M	8 K / 9 K
TT2-Xrce	En / Fr	8 K / 10 K	0.6 M / 0.7 M	10 K / 10 K
TT2-Xrce	En / Ge	7 K / 19 K	0.6 M / 0.5 M	10 K / 10 K
TT2-EU	En / Sp	84 K / 97 K	5.9 M / 6.6 M	20 K / 23 K
TT2-EU	En / Fr	85 K / 91 K	6.0 M / 6.6 M	20 K / 23 K
TT2-EU	En / Ge	87 K /153 K	6.5 M / 6.1 M	20 K / 19 K

Table 2 Examples of sentence pairs from the different corpora

MLT

Spanish: Se añade un círculo grande y oscuro muy por debajo del círculo pequeño y oscuro y del triángulo.

English: A large dark circle is added far below the small dark circle and the triangle.

EuTrans-0/I

Spanish: Deseamos reservar dos habitaciones individuales con aire acondicionado hasta el próximo viernes.

English: We want to book two single rooms with air conditioning until next Friday.

EuTrans-II

Italian: Buongiorno senta io avevo prenotato una stanza singola ma non la vorrei piu', vorrei disdire la prenotazione.

English: Good-morning listen I had reserved a single room but I would not like it anymore, I would like to cancel the reservation.

Ametra-Met

Spanish: En la primera mitad del día cielos nubosos a muy nubosos, con lloviznas, sobre todo en el litoral y áreas de montaña y más probables al este.

Basque: Egunaren lehen zatian zerua hodeitsu edo oso hodeitsu azalduko da eta zirimiria botako du, bereziki kostaldean eta mendi inguruetan eta aukera handiagoz ekialdean.

TT2-Xrce

English: Although these two types of scan jobs do not require network rights for the users, the appropriate distribution template(s) must first be created by the system a dministrator or a template user.

Spanish: A pesar de que estos dos tipos de trabajos de exploración no requieren derechos de red para los usuarios, el administrador del sistema o el usuario de la plantilla deberán crear primero las plantillas de distribución apropiadas.

German: Obwohl bei diesen beiden Auftragsarten kein e Netzwerkrechte für den Benutzer nötig sind, müssen die geeigneten Verteilerpro file erst vom Systemadministrator oder einem Profilbenutzer erstellt werden.

French: Bien que ces deux types de travaux ne nÈces sitent pas de droits d'accès pour l'utilisateur, l'administrateur système ou un utilisateur de modèle doit créer le(s) modèle(s) de distribution approprié(s).

TT2-EU

English: The threat posed by the Year 2000 problem is not restricted to particular sectors of activity, nor to any specific country or region.

Spanish: La amenaza generada por el efecto 2000 no está restringida a sectores concretos de actividad ni a ningún país o región determinados.

French: La menace que fait peser le problème du passage à l'an 2000 ne se limite pas à des secteurs d'activité particuliers, ni à des pays ou des régions spécifiques.

German: Die von dem Computerprob lem der Jahrtausendwende ausgehenden Gefahren sind nicht auf bestimmte Sektoren der Wirtschaftstätigkeit oder der Staatsverwaltung oder auf bestimmte Länder oder Regionen beschränkt.

the corresponding consortia and, upon request, they can be made available free of charge for non-commercial use.

The tasks in Table 1 are ordered according to increasing task-complexity. The complexity of a translation task has many dimensions. Perhaps the most important factors are: (a) vocabulary sizes, (b) word-order differences in the language pair and (c) scarcity of bilingual data available for training the system.

The simplest application of Table 1, first described in Castellanos, Galiano and Vidal (1994), was the *Miniature Language Translation* (MLT) task. It was a direct extension for MT of the so called *Miniature Language Acquisition* task (MLA), originally introduced for testing the capabilities of (language) learning systems (Feldman et al., 1990).

MLT was a small task, with vocabularies of about 30 words, involving the translation of sentences used to describe and manipulate simple visual scenes. Three languages were considered: Spanish, English and German. Training and testing data were automatically generated according to the task specifications (Feldman et al., 1990; Castellanos, Galiano, & Vidal, 1994). For the experiments reviewed in the next section, $8K$ Spanish-English sentence pairs (about $120K$ running words per language) were used for training. Testing was performed on another $10K$ sentences different from those used in training (Vilar et al., 1996a).

Next application in difficulty was the EUTRANS-0 task, called *traveler task* in Vidal (1997). It was a much larger and practically motivated task established in the framework of the European Union speech-to-speech MT project EUTRANS (EuTrans, 2000). The task involved human-to-human communication situations in the front-desk of a hotel. Three large parallel corpora for this task were produced in a semi-automatic way for three language pairs: Spanish-English, Spanish-German and Spanish-Italian (Amengual et al., 2000).

The Spanish-English corpus contained $500K$ ($171K$ different) sentence pairs, with Spanish/English vocabulary sizes of 689/514 words and test-set bigram perplexities of 6.8/5.6, respectively. Since the total size of this corpus was considered unrealistically large (almost $5M$ running words per language), a much reduced corpus, called EUTRANS-I, was built by randomly selecting $10K$ ($6.8K$ different) sentence pairs for training and $3K$ (all different) for testing.

The EUTRANS task has proved very useful for development purposes and, recently, three new EUTRANS-I corpora have been produced for MT and STSMT experimentation for the following pairs of languages: Spanish-Catalan, Spanish-Basque and Portuguese-English (González et al., 2002; Picó et al., 2005; Pérez et al., 2005). For the sake of brevity, only Spanish-English data and results will be discussed here.

Apart from the multilingual corpora mentioned above, an additional Italian-English MT corpus was produced in the EUTRANS project (EuTrans, 2000). This corpus, referred to as EUTRANS-II, corresponds to a task significantly more complex and closer real life than those previously considered. In this case, a speech corpus was acquired by recording real phone calls to the (simulated) front desk of a hotel. An associated text corpus was obtained by manually transcribing the acquired Italian utterances and translating them into English. The resulting corpus is much smaller than the previous ones, while having four times larger vocabularies ($2.5K/1.7K$ words). From this corpus, approximately $3K$ pairs of sentences (about $60K$ running words per language) were used for training the translation models and about 300 sentences were used for testing.

One of the most challenging pairs of languages we have ever considered is Spanish-Basque. The Basque is one of the official languages spoken in one part of Spain. The AMETRA project (González et al., 2004) was devoted to the development of tools to increase productivity of official stationery. Two specific tasks were studied, but we will only focus on one of them: AMETRA-MET. This corpus contains approximately $80K$ running words per language and consists of transcriptions of meteorological forecasts from a Basque TV channel. This corresponds to a relatively low-perplexity task with rather restricted semantic domain and vocabulary sizes around $1K$ words.

Recently, other applications have been considered in the context of the computer assisted translation (CAT) project *Trans-Type 2* (TT2) (Civera et al., 2004). Two sets of corpora were produced in this project: TT2-XRCE and TT2-EU, each set encompassing three language pairs, English-Spanish, English-French and English-German (six pairs, considering both translation directions) (Khadivi & Goutte, 2003).

The first corpora, TT2-Xrce, is a collection of technical Xerox manuals which involve relatively large vocabularies of 7–19 K words and contain around 0.6 M running words per language. The second one, TT2-EU, corresponds to translations of the Bulletin of the European Union, which are publicly available on the Internet. The corpora used in the experiments reviewed in the next section contain about 6.3 M running words per language and the vocabulary sizes range from 84 to 154 K words.

8.2. Speech data and acoustic models used for STSMT

To test the capabilities of STSMT (speech-to-speech MT) systems, several speech-input corpora associated with some of the bi-text corpora described above have been collected. The main features of these corpora are summarized below. A few details are also given about the *acoustic hidden Markov models* (HMM) (Jelinek, 1998) trained with these corpora and used in the experiments discussed in the next section.

- MLT: 120 phonetically balanced Spanish utterances from a different task, uttered by 10 speakers, were used to train 26 very simple monophone acoustic units, each modeled by a *discrete* HMM with 3 states and 128 codewords. Testing was carried out on 100 Spanish sentences, randomly selected from the text test set, uttered by a four speakers (400 utterances in total). One of the test speakers was included among the 10 training speakers (Vilar et al., 1996a).
- EUTRANS-0 and EUTRANS-I: Acoustic models of 26 Spanish monophone units were also used. In this case each model was a left-to-right *continuous density* HMM (CDHMM). These models were trained with the HTK Toolkit (Young et al., 1997) using a Spanish corpus of 57K running words from 315 speakers (*Albayzin* corpus (Díaz-Verdejo et al., 1998)). 336 utterances (3K running words) by 4 speakers, not involved in training, were used for testing. Two versions of these speech data were used: The first one had *microphone quality*; i.e., 7.8 kHz bandwidth, 16 kHz sampling rate and 16 bits per sample linear resolution). From these data a second *telephone-quality* speech data set was produced, with 3.8 kHz bandwidth, 8 kHz sampling rate and 10 bits per sample linear resolution.
- EUTRANS-II: The speech corpus consisted of 7.9 hours of speech, uttered by 276 speakers over standard telephone lines. 52K running words by 252 speakers were used to train the acoustic HMMs and 278 utterances by 24 non-training speakers (5K running words) were used for testing (Casacuberta et al., 2004). In this case, more sophisticated acoustic modeling was adopted. More specifically, decision-tree clustered generalized CDHMM triphones (CART with 1,500 tied states plus silence) were trained using the HMM training tools of the RWTH (Aachen) speech processing group (Ney et al., 1998).

8.3. Performance measures

Measuring the accuracy of MT, STSMT and CAT systems is a difficult problem. In this article only automatically computed *objective* scores will considered, thereby avoiding to discuss assessment results based on *subjective*, human-expert based judgments. Generally speaking, these objective measures correlate fairly well with the quality subjectively assessed by human experts (Nießen et al., 2000; Nießen, 2003). However, given that a correct translation of a given source utterance is by no means unique, objective measures based on single-reference test data always tend to be pessimistic.

- *Translation Word error Rate* (TWER) is the minimum number of word insertions, substitutions and deletions needed to match the system output with a *single* target sentence reference divided by the total number of reference words (Amengual et al., 2000; Nießen et al., 2000).
- *Key-Stroke Ratio* (KSR) is a measure used in the context of CAT to estimate the effort needed by a human translator to produce correct translations with the help of the CAT system. It is computed as the percentage of keys-strokes that a human translator had to type with respect to those needed to type the entire text without the help of the system (Cubel et al., 2003; Och, Zens, & Ney, 2003).

KSR is only relevant in CAT experiments and TWER applies both to STSMT and text-input MT. In real MT applications involving large vocabularies, a TWER of 20–30% is generally considered a reasonably good, useful result, and the same is true for the KSR in CAT applications.

In the MT community another interesting measure has recently been proposed, which is called *BiLingual Evaluation Understudy metric (BLEU)* (Papineni et al., 2002). Since many of the results overviewed here are older than the introduction of BLEU, for the sake of homogeneous presentation, this measure will not be reported in any of the experiments of the next section. Other measures that aimed at being closer to human evaluation have been proposed in Nießen et al. (2000) and Nießen (2003), but they were too expensive or time-consuming to be applicable in many of the experiments here considered.

9. Results

Three types of results are reported to check the feasibility of stochastic finite state transducers for machine translation: Text-input MT, STSMT and CAT. Whenever available, these results are accompanied by others obtained by the well known AT statistical approach outlined in Section 7 (Och & Ney, 2004), using the very same training and test data. In many cases, results obtained by other state-of-the-art "competitive" (i.e., not based on SFST technology) MT techniques are also available. However, for the sake of presentation clarity, only the AT results will be presented, even though, in some cases, somewhat better results were obtained using the so called *phrase-based* approach (Tomás & Casacuberta, 2001).

9.1. Text-input translation results

Here we review the most significant MT experiments carried out with the different SFST learning techniques outlined in this article as applied to some of the corpora described in the previous section. A summary of selected results is given in Table 3.

OSTIA and OSTIA-DR have been applied to many relatively simple MT tasks. The first works, reported in Castellanos, Galiano, and Vidal (1994), were carried out on the MLT task. This task provided the first testbed for studying the capabilities and shortcomings of basic SFST learning techniques. It allowed, for instance, to understand the need for source and/or target language models in order to cope with insufficient training data and/or imperfect input text. All in all, very good results (TWER below 1%) were achieved for the MLT corpus using OSTIA-DR along error-correcting smoothing (Vilar, Vidal, & Amengual, 1996; Vilar et al., 1996a).

Table 3 Summary of Translation Word Error Rate (TWER %) for selected tasks of increasing difficulty—see details in Table 1. In all the cases languages involved are Spanish-English, except for AMETRA-MET and EUTRANS-II which involve Spanish-Basque and Italian-English, respectively. Best results are typeset in boldface. AT stands for *statistical alignment templates* (see text for details)

Task	Year	OSTIA	OSTIA-DR	OMEGA	GIATI	AT
MLT	1993	3	<**1**	–	–	–
EUTRANS-0	1996	≈1	<**1**	<**1**	3	–
EUTRANS-I	1998	–	>70	**7**	**7**	–
EUTRANS-II	1999	–	>80	37	**25**	**25**
AMETRA-MET	2004	–	–	–	40	–
TT2-XRCE	2004	–	–	–	45	**40**
TT2-EU	2004	–	–	–	52	**47**

The application of OSTIA and OSTIA-DR to the more realistic EUTRANS task was first described in Vidal (1997). Very good results with EUTRANS-0 were reported in Vidal (1997) and additional results, both for EUTRANS-0 and EUTRANS-I, can be seen in Amengual et al. (2000). Using a categorized[4] version of the huge EUTRANS-0 training corpus, OSTIA-DR produced almost perfect models, with TWER lower than 1%. However, results degraded significantly when the more realistically sized EUTRANS-I training corpus was used. Since the learned models were clearly under-trained, error-correcting smoothing was needed in this case, leading to a text-input TWER close to 10% (EuTrans, 2000). Without using categories the TWER was exceedingly high (>70%) to be considered useful.

OMEGA was also tested on the EUTRANS-0 and EUTRANS-I corpora. While results on EUTRANS-0 were similar to or slightly worse than those of OSTIA, on the much smaller EUTRANS-I corpus, OMEGA was clearly better. Without using categories, OMEGA achieved error-correcting text-input TWER better than 7% (EuTrans, 2000) whereas, under the same conditions, OSTIA(-DR) completely failed to produce useful results.

Regarding the more difficult EUTRANS-II task, OSTIA could not be used at all in this task and OMEGA produced only moderately acceptable results (37% TWER) (Casacuberta et al., 2004).

The GIATI technique was tested with all the corpora of the EUTRANS project. TWER smaller than 7% and 25% were obtained for text input with EUTRANS-I and EUTRANS-II, respectively (Casacuberta & Vidal, 2004). The EUTRANS-II results are clearly better than those achieved by OMEGA under the same conditions. Overall, GIATI was among the best techniques tested in the framework of the EUTRANS project (EuTrans, 2000).

GIATI was also extensively used in the TT2 project (Barrachina et al., 2006). Apart from the CAT experiments which will be discussed later, plain MT empirical tests were also carried out using the TT2 corpora involving three language pairs (six by considering both translation directions of each pair). Table 3 shows only the Spanish-to-English results. For the TT2-XRCE and TT2-EU corpora, GIATI achieved 45% and 52% TWER, respectively, which were overcome by other statistical approaches. Results for the other languages pairs roughly showed a similar tendency.

GIATI has recently been successfully applied to versions of the EUTRANS-I task for Catalan-Spanish (González et al., 2002), Portuguese-English (Picó et al., 2005) and

[4] Seven categories: proper names, dates, times of day, etc. Each instance of these categories was substituted by a corresponding non-terminal symbol.

Spanish-Basque (Pérez et al., 2005). Also in the Spanish-to-Basque pair, GIATI has been applied to the more practical, real task AMETRA-MET, where it achieved 40% TWER (González et al., 2004).

9.2. Speech-input translation results

In all the experiments discussed in this subsection, source-language utterances are translated into target-language sentences. These (text) sentences can be easily and accurately converted into target-language utterances by using available *text-to-speech* synthesizers.

The relative quality of the translations obtained by using the *serial* or the *integrated* architectures is discussed in Casacuberta et al. (2004). Here these details are omitted and, for each translation task and technique, only the best results are discussed (a summary appears in Table 4).

The first *speech-input* experiments were carried out with OSTIA-DR applied to the MLT task, leading to very good performance (3% TWER) with very simple (*discrete*) acoustic models (Vidal, 1997). Using better (*continuous-density*) acoustic modeling, OSTIA-DR also achieved very good results in (microphone) speech-input experiments with EUTRANS-0 (Vidal, 1997; Amengual et al., 2000).

OMEGA transducers learned with EUTRANS-0 and EUTRANS-I were also tested with speech-input. Results were good for EUTRANS-0 and only moderately good for EUTRANS-I, with about 13% and 18% TWER for microphone and telephone speech, respectively (Casacuberta et al., 2004). Regarding the significantly more difficult EUTRANS-II task, OSTIA could not be used at all and OMEGA produced only moderately acceptable results: less than 50% TWER (Casacuberta et al., 2004).

The GIATI technique was successfully tested with all the speech corpora of the EUTRANS project. While results for EUTRANS-0 were slightly worse than those of OSTIA-DR or OMEGA, GIATI clearly overcame these techniques when trained with the more realistic EUTRANS-I and EUTRANS-II corpora. In EUTRANS-I, GIATI achieved 8% and 13% TWER for microphone and telephone speech-input, respectively, and in the (telephone-input) experiments with EUTRANS-II it produced about 38% TWER. Overall, GIATI was among the best techniques tested in the framework of the EUTRANS project (EuTrans, 2000).

Further GIATI speech-input translation results for Catalan-to-Spanish, Portuguese-to-English and Spanish-to-Basque tasks can be found in González et al. (2002), Picó et al. (2005) and Pérez et al. (2005).

Table 4 Speech-input *word error rate* and *translation word error rate* (TWER %) for different STSMT tasks. AT stands for *statistical alignment templates*

Task	Acoustics	OSTIA-DR	OMEGA	GIATI	AT
MLT	microphone	**3**	–	–	–
EUTRANS-0	microphone	**2**	4	6	5
	telephone	–	**8**	11	10
EUTRANS-I	microphone	–	13	8	7
	telephone	–	18	**13**	**13**
EUTRANS-II	telephone	–	49	**38**	**38**

Table 5 Key Stroke Ratio (KSR %) for two CAT tasks in different languages. AT stands for *statistical alignment templates*. Languages: En = English, Sp = Spanish, Fr = French and Ge = German

Task	Languages	GIATI	AT
XRCE	En - Sp	**22**	23
	Sp - En	27	**24**
	En - Fr	44	**40**
	Ge - En	47	**46**
EU	En - Sp	33	33
	Sp - En	31	31
	En - Fr	**30**	34
	Ge - En	39	**38**

9.3. Computer-assisted translation results

Recently, GIATI has also been used in the context of the CAT project *Trans-Type 2* (TT2) (Civera et al., 2004; Cubel et al., 2005). Results on the TT2-XRCE task were promising and were even better for most language pairs of the more difficult TT2-EU task (Table 5). As with the results reported in Section 9.1, in these tasks of moderate/large difficulty, the CAT performance of GIATI also tends to fall behind that of other statistical approaches. Further results for French to English and English to German on TT2-XRCE and TT2-EU can be found in Cubel et al. (2005).

Apart from these results, GIATI has been used in one of the prototypes which have shown better practical behavior according to real human tests on the XEROX task (Macklovitch, 2004).

10. Conclusions

A number of techniques to learn stochastic finite-state transducers for machine translation have been reviewed. One of the interesting reasons for the use of SFST in MT is the existence of very efficient search algorithms for translation. On the other hand, finite-state transducer inference techniques are generally less computationally expensive than most pure statistical approaches. Although it can be argued that SFST are too simplistic models to cope with the complex nature of natural language, it is also well known that many natural language phenomena can be adequately represented by finite-state devices (Kornai, 1985). Finally, SFST allow for an easy integration with other knowledge sources, such as acoustics models, leading to very effective solutions to speech translation.

Our review has started with techniques mainly falling under the traditional grammatical inference paradigm. While these techniques have proven able to learn very adequate MT models for non-trivial tasks, the amount of training sentence pairs needed often becomes prohibitive in real-world situations. Other techniques have also been reviewed that circumvent this problem by increasingly relying on statistically-derived information.

As task complexity increases, we think that statistically-based learning is the most promising framework, particularly in those tasks where the assumption of left-to-right alignments leads to huge SFSTs. By explicitly allowing non left-to-right alignments, statistical models can represent more complex natural language phenomena in a more compact form, but they generally exhibit very high computational learning and translation costs as compared with the corresponding algorithms for finite-state models. This problem has been addressed in Vilar, Vidal, and Amengual (1996), where a *word-reordering* of the target sentence, based on statistical alignments, is proposed for SFST learning. The reordering algorithm adds some meta-symbols to the resulting target sentences that can later help recover the correct word order. While results presented in Vilar, Vidal, and Amengual (1996) for the (very simple) MLT task were promising, further studies revealed that more work is actually needed to develop adequate word/phrase reordering techniques that would lead to significant SFST size reduction in real MT applications. Other promising techniques proposed to reduce the size of SFST trained with large corpora are based on word categorization (Barrachina & Vilar, 1999). However, again, further studies are needed to assess the possibilities of these techniques to achieve significant size reductions, without losing translation accuracy.

With respect to the different SFST learning techniques tested, the results presented in this article clearly suggest that GIATI is the only one that can cope with translation tasks under real conditions of vocabulary sizes and amounts of training data available. However, as task complexity increases, GIATI tends to fall behind other approaches that rely more explicitly on statistics.

One of the shortcomings of GIATI comes from the fact that it needs a *given* bilingual segmentation of the training pairs in order to determine the *"extended alphabet"* and the corresponding morphisms. Clearly, for a given translation pair, there are many possible bilingual segmentations; but the present version of GIATI does not take advantage of this fact, thereby making less profit from the (always scarce) training data. A possible way to deal with this problem is to develop a new GIATI version which is more explicitly based on statistics, as proposed in Vidal and Casacuberta (2004) and Andrés (2005). The main idea is to consider the bilingual segmentation of the training pairs as a "hidden variable". The EM algorithm is then used to simultaneously optimize the segment boundaries and the n-gram parameters that model the concatenation of the bilingual segments. Therefore, an additional advantage of this derivation is that it no longer has to rely on "external" statistical techniques to obtain an "adequate bilingual segmentation", or make use of heuristics to transform pairs of sentences into conventional strings. Finally, it should be emphasized that this new GIATI version can retain the interesting advantage that the learned models directly admit a simple finite-state representation.

All the reviewed techniques have been tested on practical MT tasks considered in many Spanish and European projects and company contracts, involving a large variety of languages such as Spanish, English, Italian, German, French, Portuguese, Catalan and Basque. As a result of these projects a number of prototypes have been implemented and successfully tested under real (or at least realistic) conditions. On-line demonstrations of some of these prototypes are available at `http://prhlt.iti.es/demos/demos.htm`.

Acknowledgments The authors wish to thank the anonymous reviewers for their criticisms and suggestions.

References

Alshawi, H., Bangalore, S., & Douglas, S. (2000). Learning dependency translation models as collections of finite state head transducers. *Computational Linguistics*, *26*(1), 45–60.

Amengual, J., Benedí, J., Casacuberta, F., Castaño, A., Castellanos, A., Jiménez, V., Llorens, D., Marzal, A., Pastor, M., Prat, F., Vidal, E., & Vilar, J. (2000). The EuTrans-I speech translation system. *Machine Translation*, *15*, 75–103.

Amengual, J., Sanchis, A., Vidal, E., & Benedí, J. (2001). Language simplification through error-correcting and grammatical inference techniques. *Machine Learning*, *44*, 143–159.

Amengual, J., & Vidal, E. (1998). Efficient error-corecting viterbi parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(10), 1109–1116.

Andrés, J. (2005). N-HSEST: N-history segmented enumerable stochastic transducer. Technical Report DSIC-II/16/05, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia.

Bangalore, S., & Riccardi, G. (2003). Stochastic finite-state models for spoken language machine translation. *Machine Translation*, *17*(3), 165–184.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., & Vilar, J.-M. (2006). Statistical approaches to computer-assisted translation. *To be published*.

Barrachina, S., & Vilar, J. M. (1999). Bilingual clustering using monolingual algorithms. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99)*, (pp. 77–87), University College, Chester, England.

Berger, A. L., Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., Gillett, J. R., Kehler, A. S., & Mercer, R. L. (1996). Language translation apparatus and method of using context-based translation models. United States Patent, No. 5510981.

Berstel, J. (1979). *Transductions and context-free languages*. B. G. Teubner Stuttgart.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roosin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, *16*(2), 79–85.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263–311.

Casacuberta, F. (2000). Inference of finite-state transducers by using regular grammars and morphisms. In *Grammatical inference: Algorithms and applications. Proceedings of the 5th ICGI*, vol. 1891 of *Lecture Notes in Computer Science* (pp. 1–14). Springer-Verlag.

Casacuberta, F., & de la Higuera, C. (2000). Computational complexity of problems on probabilistic grammars and transducers. In *Grammatical inference: Algorithms and applications. Proceedings of the 5th ICGI*, vol. 1891 of *Lecture Notes in Computer Science* (pp. 15–24). Springer-Verlag.

Casacuberta, F., Ney, H., Och, F. J., Vidal, E., Vilar, J. M., Barrachina, S., García-Varea, I., Llorens, D., Martínez, C., Molau, S., Nevado, F., Pastor, M., Picó, D., Sanchis, A., & Tillmann, C. (2004). Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*, *18*, 25–47.

Casacuberta, F., & Vidal, E. (2004). Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, *30*(2), 205–225.

Casacuberta, F., Vidal, E., & Picó, D. (2005). Inference of finite-state transducers from regular languages. *Pattern Recognition*, *38*, 1431–1443.

Castellanos, A., Galiano, I., & Vidal, E. (1994). Application of OSTIA to machine translation tasks. In R. C. Carrasco, & J. Oncina (Eds.), *Grammatical inference and applications, Proceedings of 2nd ICGI*, vol. 862 of *Lecture Notes in Computer Science* (pp. 93–105). Springer-Verlag.

Castellanos, A., Vidal, E., Varó, A., & Oncina, J. (1998). Language understanding and subsequential transducer learning. *Computer Speech and Language*, *12*, 193–228.

Charniak, E., Knight, K., & Yamada, K. (2003). Syntax-based language models for statistical machine translation. In *Proceedings of the Machine Translation Summit IX*, (pp. 40–46), New Orleans, Louisiana, USA.

Civera, J., Vilar, J., Cubel, E., Lagarda, A., Barrachina, S., Casacuberta, F., Vidal, E., Picó, D., & González, J. (2004). A syntactic pattern recognition approach to computer assisted translation. In A. Fred, T. Caelli, A. Campilho, R. P. Duin, & de Ridder, D. (Eds.), *Advances in statistical, structural and syntactical pattern recognition. Proceedings of the Joint IAPR International Workshops on Syntactical and Structural Pattern Recognition and Statistical Pattern Recognition*, Lecture Notes in Computer Science (pp. 207–215). Springer-Verlag.

Cubel, E., Civera, J. F. C., Lagarda, A. L., Vilar, J. M., & Vidal, E. (2005). Hybrid stochastic finite-state learning techniques and search algorithms (deliverable d5.3). Technical report, TransType2(IST-2001-32091), ITI.

Cubel, E., González, J., Lagarda, A., Casacuberta, F., Juan, A., & Vidal, E. (2003). Adapting finite-state translation to the transtype2 project. In *Proceedings of the Joint Conference Combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop* (pp. 54–60), Dublin, Ireland.

Díaz-Verdejo, J., Peinado, A., Rubio, A., Segarra, E., Prieto, N., & Casacuberta, F. (1998). Albayzin: A task oriented spanish speech corpus. In *Proceedings of First International Conference on Language Resources and Evaluation (LREC-98)* (vol. 1, pp. 497–501).

EuTrans (2000). Example-based language translation systems. Final report. Technical report, Instituto Tecnológico de Informática, Fondazione Ugo Bordoni, Rheinisch Westfälische Technische Hochschule Aachen Lehrstuhl für Informatik VI, Zeres GmbH Bochum: Long Term Research Domain, Project Number 30268.

Extra (1997). Extensions of the text and speech translation system to restricted domains that can be learnt from examples. Universitat Jaume I and Universitat Politècnica de València, CICYT project.

Feldman, J. A., Lakoff, G., Stolcke, A., & san Hollbach Weber, S. (1990). Miniature Language Acquisition: A touchstone for cognitive science. Technical Report TR-90-009, International Computer Science Institute, Berkeley, California (EEUU).

Germann, U. (2003). Greedy decoding for statistical machine translation in almost linear time. In M. Hearst, & M. Ostendorf (Eds.), *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)* (pp. 72–79), Edmonton, Alberta, Canada.

González, J., Nevado, F., Navarro, J., Pastor, M., Casacuberta, F., Vidal, E., Fabregat, F., de Val, J., Arenas, L., Pla, F., & Tomás, J. (2002). SISHITRA: Sistemas de traducción catalán-castellano y castellano-catalán con entrada de texto y voz. In *Actas de las II Jornadas en Tecnología del Habla*, Granada. Red Temática en Tecnologías del Habla.

González, J., Ortiz, D., Tomás, J., & Casacuberta, F. (2004). A comparison of different statistical machine translation approaches for Spanish-to-Basque translation. In *Proceedings of the Terceras Jornadas de Tecnología del Habla* (pp. 213–218). Valencia, Spain.

Jelinek, F. (1998). *Statistical methods for speech recognition*. Cambridge, MA. The MIT Press.

Khadivi, S. & Goutte, C. (2003). Tools for corpus alignment and evaluation of the alignments (deliverable d4.9). Technical report, TransType2(IST-2001-32091), RWTH Aachen and Xerox Co.

Knight, K. (1999). Decoding complexity in word-replacement translation models. *Computational Linguistics*, *25*(4), 607–615.

Knight, K., & Al-Onaizan, Y. (1998). Translation with finite-state devices. In *Proceedings of the 4th AMTA Conference* (pp. 421–437). Langhorne, Pennsylvania.

Knight, K., Al-Onaizan, Y., Curin, J., Jahr, M., Lafferty, J., Melamed, D., Purdyand, D., Och, F., Smith, N., & Yarowsky, D. (1999). EGYPT a statistical machine translation toolki. http://www.clsp.jhu.edu/ws99/projects/mt/. An NFS Workshop: Language Engineering for Students and Professionals Integrating Research and Education.

Koehn, P. (2004). Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the The 6th Conference of the Association for Machine Translation in the Americas (AMTA04)*, vol. 3265 of *Lecture Notes in Artificial Intelligence* (pp. 115–124). Georgetown University, Washington DC, USA: Springer.

Kornai, A. (1985). Natural languages and the chomsky hierarchy. In *Proceedings of the Second Conference on European CChapter of the Association for Computational Linguistics* (pp. 1–7). Morristown, NJ, USA. Association for Computational Linguistics.

Kumar, S., & Byrne, W. (2003). A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 63–70). Edmonton.

Kumar, S., Deng, Y., & Byrne, W. (2006). A weighted finite state transducer translation template model for statistical machine translation. *Natural Language Engineering 12*(1), 35–75.

Langlais, P., Foster, G., & Lapalme, G. (2000). TransType: A computer-aided translation typing system. In *Proceedings of the Workshop on Embedded Machine Translation Systems (NAACL/ANLP2000)* (pp. 46–52). Seattle, Washington.

Llorens, D., Vilar, J. M., & Casacuberta, F. (2002). Finite state language models smoothed using n-grams. *International Journal of Pattern Recognition and Artificial Intelligence*, *16*(3), 275–289.

Macklovitch, E. (2004). The contribution of end-users to the transtype2 project. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)* (pp. 197–207). TT2.

Mäkinen, E. (1999). Inferring finite transducers. Technical Report A-1999-3. University of Tampere.

Marcu, D., & Wong, W. (2002). A phrase-based joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora EMNLP-02* (pp. 133–139). Philadelphia PA, USA.

Ney, H., Martin, S., & Wessel, F. (1997). Statistical language modeling using leaving-one-out. In S. Young, & G. Bloothooft (Eds.), *Corpus-Based Statiscal Methods in Speech and Language Processing* (pp. 174–207). Kluwer Academic Publishers.

Ney, H., Nießen, S., Och, F. J., Sawaf, H., Tillmann, C., & Vogel, S. (2000). Algorithms for statistical translation of spoken language. *IEEE Transactions on Speech and Audio Processing*, *8*(1), 24–36.

Ney, H., Welling, L., Ortmanns, S., Beulen, K., & Wessel, F. (1998). The RWTH large vocabulary continuous speech recognition system. In *Proceeding of the International Conference on Acoustic, Speech and Signal Processing* (pp. 853–856).

Nießen, S. (2003). Improving statistical machine translation using morpho-syntactic information. PhD thesis, RWTH Aachen, Germany.

Nießen, S., Och, F., Leusch, G., & Ney, H. (2000). An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (pp. 39–45). Athens, Greece.

Och, F., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

Och, F., & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, *30*(4), 417–449.

Och, F., Zens, R., & Ney, H. (2003). Efficient search for interactive statistical machine translation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 387–393). Budapest, Hungary.

Oncina, J., García, P., & Vidal, E. (1993). Learning subsequential transducers for pattern recognition interpretation tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(5), 448–458.

Oncina, J., & Varó, M. (1996). Using domain information during the learning of a subsequential transducer. In *Grammatical Inference: Learning Syntax from Sentences*, vol. 1147 of *Lecture Notes on Computer Science* (pp. 313–325).

Ortiz, D., García-Varea, I., & Casacuberta, F. (2005). Thot: a toolkit to train phrase-based statistical translation models. In *Proceedings of the Tenth Machine Translation Summit* (pp. 141–148). Phuket, Thailand.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association Computational Linguistics (ACL)* (pp. 311–318) Philadelphia PA, USA.

Pérez, A., Casacuberta, F., Torres, I., & Guijarrubia, V. (2005). Finite state transducers based on K-TSS grammars for speech translation. In *Proceedings of the Finite-State Methods and Natural Language Processing (FSMNLP 2005)* (pp. 270–272) Helsinki, Finland.

Picó, D., & Casacuberta, F. (2001). Some statistical-estimation methods for stochastic finite-state transducers. *Machine Learning*, *44*, 121–141.

Picó, D., Caseiro, D., González, J., Trancoso, I., & Casacuberta, F. (2005). Finite-state transducer inference for a speech-input portuguese-to-english machine translation system. In *Proceedings of Interspeech'2005 - Eurospeech: 9th European Conference on Speech Communication and Technology* (pp. 2277–2280). Lisbon, Portugal.

Picó, D., Tomás, J., & Casacuberta, F. (2004). GIATI: A general methodology for finite-state translation using alignments. In *Statistical, Structural and Syntactical Pattern Recognition. Proceedings of the Joint IAPR Intternational Workshops SSPR2004 and SPR2004*, vol. 3138 of *Lecture Notes in Computer Science* (pp. 216–223). Lisboa, Portugal: Springer-Verlag.

SisHiTra (2001). Hybrid systems for the catalan-to-spanish transaltion from text and speech. Instituto Tecnológico de Informática and Universitat d'Alacant, CICYT project.

TeFaTe (2003). Finite-state technologies for specialized translation. Universitat Politècnica de València and Universitat d'Alacant, CICYT project.

Tillmann, C., & Ney, H. (2003). Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, *29*(1), 97–133.

Tomás, J., & Casacuberta, F. (2001). Monotone statistical translation using word groups. In *Proceedings of the Machine Translation Summit VIII* (pp. 357–361). Santiago de Compostela.

Tomás, J., & Casacuberta, F. (2006). Monotone and non-monotone phrase-based statistical machine translation. *To be published*.

Tomás, J., Lloret, J., & Casacuberta, F. (2005). Phrase-based alignment models for statistical machine translation. In *Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis*, vol. 3523 of *Lecture Notes in Computer Science* (pp. 605–613). Estoril (Portugal), Springer-Verlag.

Tracom (1995). Spoken language translation and understanding using learning techniques from examples. Universitat Politècnica de València and Universitat Jaume I, CICYT project.

TransType2 (2001). TT2. TransType2—Computer Assisted Translation. Project Technical Annex.

Tsukada, H., & Nagata, M. (2004). Efficient decoding for statistical machine translation with a fully expanded WFST model. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 427–433). Barcelona.

Vidal, E. (1997). Finite-state speech-to-speech translation. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP-97),* (vol.1, pp. 111–114). Munich.

Vidal, E., & Casacuberta, F. (2004). Learning finite-state models for machine translation. In *Grammatical inference: Algorithms and applications. Proceedings of the 7th International Coloquium ICGI 2004*, vol. 3264 of *Lecture Notes in Artificial Intelligence* (pp. 16–27). Athens, Greece: Springer, Invited conference.

Vidal, E., García, P., & Segarra, E. (1989). Inductive learning of finite-state transducers for the interpretation of unidimensional objects. In R. Mohr, T. Pavlidis, & A. Sanfeliu (Eds.), *Structural pattern analysis* (pp. 17–35). World Scientific pub.

Vidal, E., Thollard, F., de la Higuera, F. C. C., & Carrasco, R. (2005). Probabilistic finite-state machines—part II. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(7), 1026–1039.

Vilar, J. M. (2000). Improve the learning of subsequential transducers by using alignments and dictionaries. In *Grammatical Inference: Algorithms and Applications*, vol. 1891 of *Lecture Notes in Artificial Intelligence* (pp. 298–312). Springer-Verlag.

Vilar, J. M., Jiménez, V. M., Amengual, J. C., Castellanos, A., Llorens, D., & Vidal, E. (1996a). Text and speech translation by means of subsequential transducers. *Natural Language Engineeering*, *2*(4), 351–354. Special issue on Extended Finite Models of Language.

Vilar, J. M., Vidal, E., & Amengual, J. C. (1996b). Learning extended finite-state models for language translation. In *Proc. of Extended Finite State Models Workshop (of ECAI'96)* (pp. 92–96). Budapest.

Wu, D. (1995). Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In *Proceedings of 14th International Joint Conference on Artificial Intelligence* (pp. 1328–1335). Montreal, Canada.

Young, S., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (1997). *The HTK Book (Version 2.1)*. Cambridge University Department and Entropic Research Laboratories Inc.

Zens, R., & Ney, H. (2004). Improvements in phrase-based statistical machine translation. In *Proceedings of the Human Language Technology Conference (HLT-NAACL)* (pp. 257–264). Boston, MA.

Zens, R., Och, F., & Ney, H. (2002). Phrase-based statistical machine translation. In G. L. M., Jarke, & J. Koehler, (Eds.), *Advances in artificial intelligence. Proceedings of the 25. Annual German Conference on AI, KI 2002*, vol. 2479 of *LNAI* (pp. 18–32). Springer Verlag.