A comparison of linguistically and statistically enhanced models for speech-to-speech machine translation

Alicia Pérez[(1)], *Víctor Guijarrubia*[(1)], *Raquel Justo*[(1)], *M. Inés Torres*[(1)], *Francisco Casacuberta*[(2)]

[(1)]Dep. Electricity and Electronics. University of the Basque Country
manes@we.lc.ehu.es
[(2)]Dep. of Information Systems and Computation. Technical University of Valencia
fcn@dsic.upv.es

Abstract

The goal of this work is to improve current translation models by taking into account additional knowledge sources such as semantically motivated segmentation or statistical categorization. Specifically, two different approaches are discussed. On the one hand, phrase-based approach, and on the other hand, categorization. For both approaches, both statistical and linguistic alternatives are explored. As for translation framework, finite-state transducers are considered. These are versatile models that can be easily integrated on-the-fly with acoustic models for speech translation purposes. In what the experimental framework concerns, all the models presented were evaluated and compared taking confidence intervals into account.

1. Introduction

Stochastic finite-state transducers (SFST) [1] consist of a sub-set of probabilistic translation models [2]. They are versatile models that count on efficient algorithms for *inference* from training samples [3], *composition* with other finite-state models [4, 5] that allows for a hierarchical structure of several knowledge sources, *minimization* and *decoding* [6].

The main goal of this paper is to carry out a comparative study on the benefits of additional knowledge sources within finite-state framework constrained to GIATI methodology [7]. We aim at comparing the classical GIATI approach, where the source sentence to be translated is analyzed word-by-word, with category-based and phrase-based approaches, where either the analysis of the source sentence or the generation of target sentence is driven by other kind of tokens. We intend to find out wether it is worth or not to create more complex models such as class-based or phrase-based models within this finite-state framework. In previous works we studied such models taking just linguistic approach into account. The contribution of this paper is to tackle those models under purely statistical framework and to compare with previous models, which joined linguistic knowledge sources within the statistical framework. The experimental layout entails both text and speech translation.

On the other hand, this work is an attempt to quantitatively determine the contribution of linguistic knowledge sources in contrast to statistical ones for the same task and under exactly the same conditions. Nevertheless, when the discrepancy in score of two systems is marginal we may wonder whether it is possible to assert that one system outperforms the other. The arising question is what a marginal difference stands for or what might be considered as a significant difference. In this context, the scores are presented along with their confidence interval. In addition, whenever we say that a system outperforms another, the results are accompanied by the *probability of improvement*.

The organization of this paper is as follows: first of all, an overview of both the baseline (in this paper also referred to as word-based), category-based and phrase-based finite-state transducers is presented in sections 2, 3 and 4 respectively. In order to quantitatively assess the performance of each model, section 5 is devoted to a thorough discussion of the experiments. Finally, section 6 summarizes the conclusions of the present work as well as the proposed lines for future investigation in this field.

2. Source words driven finite-state transducers

An SFST is a finite-state machine which accepts strings belonging to a source vocabulary and gives as a result strings belonging to a target vocabulary along with the joint probability of both source and target strings (for a formal definition turn to [1]). The characteristics defining the SFST are its topology and the probability distributions over the transitions and the states. These distinctive features can be automatically learnt from bilingual samples by efficient algorithms such as GIATI (Grammar Inference and Alignments for Transducers Inference) [7], which constitutes the specific object of study in this work. Regarding the topology, a smoothed k-Testable in the Strict-Sense (k-TSS) grammar [8, 9] (with k=3) was used for all the experiments (even for the language model in isolate speech recognition experiments). In short, k-TSS models are considered to be the syntactic approach of the well known n-gram models.

Our goal is to carry out speech translation with the mentioned SFSTs. Let us summarize how statistical speech translation works. The goal of statistical speech translation is to find the most likely translation ($\widehat{\mathbf{t}}$) given the the acoustic representation $\mathbf{x}$ of a speech signal in the source language:

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{x}) \tag{1}$$

The transcription of the speech into text is an unknown variable, $\mathbf{s}$, which might be introduced as a hidden variable.

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t}, \mathbf{s}|\mathbf{x}) \tag{2}$$

Applying the Bayes' decision rule:

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t}, \mathbf{s}) P(\mathbf{x}|\mathbf{t}, \mathbf{s}) \tag{3}$$

Let us assume that the acoustic representation of a speech signal only depends on its transcription in the source language, that is, the pronunciation of an utterance does not depend on the translation in other language. Hence, eq. (3) can be rewritten as:

$$\widehat{\mathbf{t}} = \arg\max_{\mathbf{t}} \sum_{\mathbf{s}} P(\mathbf{t}, \mathbf{s}) P(\mathbf{x}|\mathbf{s}) \tag{4}$$

There are two terms involved in eq. (4). On the one hand, $P(\mathbf{x}|\mathbf{s})$ links the acoustics with its orthographic representation. It is, somehow, a lexical model similar to those used in speech recognition. Specifically, given that the source string is formed by a concatenation of several words $\mathbf{s} = s_1^I$, having their acoustic representation $\mathbf{x} = x_1^I$, we assume the pronunciation of the word s_j to be independent of other words.

$$P(\mathbf{x}|\mathbf{s}) \simeq \prod_{i}^{I} P(x_i|s_i) \tag{5}$$

On the other hand, $P(\mathbf{t}, \mathbf{s})$ represents the probability of $(\mathbf{t}, \mathbf{s})$ to be a translation pair. The joint probability translation model might be modeled with an SFST (τ), $P(\mathbf{t}, \mathbf{s}) \simeq P_\tau(\mathbf{t}, \mathbf{s})$. Such a model might be used for either speech translation (as shown before) or by itself for text translation.

Example 1: Let us show how GIATI carries out the inference of an SFST given a couple of bilingual training samples:

$s_1 s_2 s_3 \leftrightarrow t_1 t_2 t_3$ $\qquad$ $s_1 s_2 s_4 \leftrightarrow t_1 t_2 t_4$

- *Obtain the alignments.* Let us note that in this work the involved statistical alignments were obtained with GIZA++ free toolkit [10]. Assume that for the given bilingual training set, the alignments drawn in Figure 1 were obtained (despite the fact that GIZA++ would not have given these pathologic alignments as a result).

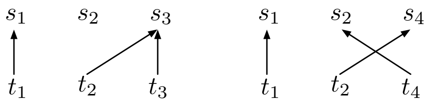


Figure 1: Alignments.

- *Get a monotonic bilingual segmentation.* On the basis of those alignments, zero or more target words are assigned to each source word, in such a way that a monotonic bilingual segmentation is obtained. Monotonicity keeps the word order of both the source and the target strings.
  $(s_1, t_1)(s_2, \lambda)(s_3, t_2 t_3)$ $\qquad$ $(s_1, t_1)(s_2, \lambda)(s_4, t_2 t_4)$
  Note that the restriction on this baseline model is that the segmentation is driven by the source words. That is, each segmentation will produce as many segments as the length of the source string. Empty words (denoted by λ) or word sequences are allowed as target segments. From now onwards we will also refer to this baseline approach as *word-based model*, even though it is not strictly word based (given that the target tokens may consist of phrases).

- *Infer a finite-state model.* The segmentation converts each training pair into a single string of an extended vocabulary composed by pairs of a source word along with a target phrase. Then from those extended strings, a regular grammar can be inferred, and thus a finite-state automaton. Note, however, that the symbols on that finite-state machine are bilingual and can be considered as input/output tokens, leading to the required transducer drawn in Figure 2.

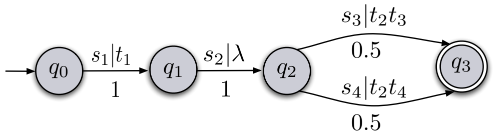


Figure 2: SFST

3. Category-based finite-state transducers

In the framework of statistical language processing large amounts of training data are required to get a robust estimation of the parameters defining the models. Data sparseness is, therefore, a problem that must be faced. One of the ways to deal with this problem is to cluster the vocabulary of the application into equivalence classes. In this way, class-based models can be used in language modeling, which is, in essence, the problem that we are tackling by means of finite-state transducers. A class-based language model is more compact and generalizes better on unseen events. Nevertheless, it only captures the relations between the categories of words while it assumes that the inter-word transition probability depends only on the word classes. As a result, it is less accurate in predicting the next word.

The category-based SFST in this work tackles the translation problem in two steps with two SFSTs: the first one has as input the source language and the categorized target as output; the second SFST takes the categorized string and converts it into a string of words in the target language as illustrated in Figure 3.

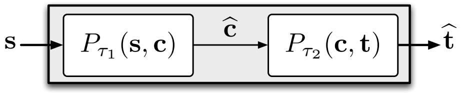

$\mathbf{s}$
$P_{\mathcal{T}_1}(\mathbf{s}, \mathbf{c})$
$\widehat{\mathbf{c}}$
$P_{\mathcal{T}_2}(\mathbf{c}, \mathbf{t})$
$\widehat{\mathbf{t}}$

Figure 3: Category-based approach.

In order to avoid the loss of information associated with the use of classical class-based models, some authors have proposed alternative approaches in speech recognition such as interpolation between word and class-based grammars [11, 12]. In our case, for speech translation the SFST itself entails a relation between classes and words. There are related works including categorization for speech translation within finite-state framework [13, 14] which significantly differ from our approach (for further details on this approach turn to [15]).

Regarding the nature of the categories, they might be either linguistically or statistically motivated, and the aim of this work is to determine if it makes a difference to use one or the other within a specific task and corpus. On the following we give some details on each categorization technique used in this work.

3.1. Linguistically motivated categories

As for linguistic categories, many criteria could have been selected, such as POS-tagging, distinguishing nouns, verbs, adjectives, gender, number, etc. In this case, lemmatization is explored as classification standard. That is, all the words sharing the same lemma are gathered within an equivalence-class. This criterion was selected since the target language involved in the task we are exploring is Basque, a highly inflected language in both nouns and verbs.

As a consequence, all the words differing only in the declension case would be grouped. Therefore, all the words in the same class share the lemma, which is in fact the main contribution in what comes to rendering the meaning. For the task under consideration (latter commented in section 5) there were 1,135 running words in Basque and they were classified in 561 classes (as they were found 561 different lemmas). Let us not that the lemmatization was carried out by Ametzagaña group[1] (a non-profit organization working on I+D) since there are not still free parsing toolkits for Basque.

3.2. Statistically motivated categories

A set of 561 statistical classes was automatically obtained by means of *mkcls* [16], a free toolkit designed to train word classes on the basis of a maximum likelihood criterion. The number of classes was set to be 561 for comparison purposes with the linguistic approach previously described.

In spite of the fact that classes were not generated using any linguistic or semantic information, in many cases, the words belonging to the same class have similarities regarding their morphologic or semantic role as shown in the following example.

Example 2: some of these statistically motivated categories of the task under consideration.

class-1:	arinduko, bihurtuko, finkatuko, handituko, helduko, nabarituko, pasatuko.
class-2:	orduetara, orduetaraino, orduetarako, ordutan.
class-3:	11, 12, 13, 15, 16, 17, 18, 19, 24, 26, 27.
class-4:	1500, 1600, 1700.
class-5:	*goradakada*, igoera.

All the words gathered in class-1 are verbs ("arindu", "bihurtu", ...) and all of them have a suffix ("-ko") representing the future tense. The class-2 assembles different declension cases of the same stem ("ordu"). Class-3 brings together numbers related to temperature, while the numbers in class-4 are related to the snow level. Nevertheless, not all the numbers within the same function are in the same category, that is, there are other categories containing numbers as well. Another odd issue is that some words that were originally incorrect due to misprints were gathered in a logical way. For instance, the word "*goradakada*" belonging to class-5 has a typographical error, it should be "gorakada", which is indeed a synonym of the word "igoera" (the other word within the class-5).

4. Phrase-based finite-state transducers

The phrase-based SFST model under GIATI approach [17] differs from existing phrase-based approaches previously defined for finite-state framework [18, 19] since the former consists of a single model that copes with both meaning transference and word reordering, while the later models entail the composition of several constituent transducers, working in separate decoding steps.

In brief, the phrase-based model proposed here is inferred from a segmented corpus considering each segment as translation unit, instead of the word. This is the main difference between phrase-based and word based approach in this work. There is no-restriction on the size of the segments. For instance, Figure 4 shows the phrase-based approach for the *Example 1* in section 2.

The inference algorithm remains unchanged from word-based to phrase-based approach, leading to a rough transducer where each transition is labeled with a phrase in the source language and a phrase in the target language (that might be empty) and a probability of that transition to occur

[1] http://ametza.com

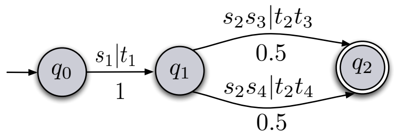
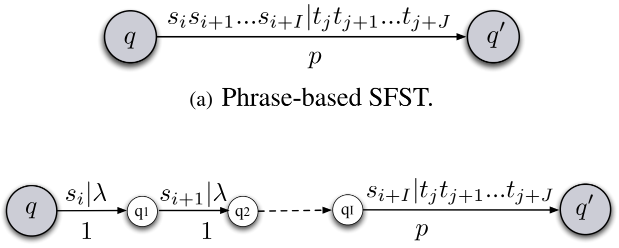


Figure 4: Phrase-based SFST.

(depicted in figure 5(a)). Each state $q \in Q$ of such a transducer represents a memory which stores the previous events, that is, the history in terms of phrases of both source and target languages. Hence, a state q is reached only after both having analyzed a specific n-gram of phrases in the source language and having produced a specific n-gram of phrases in the target language as output. The topology of the transducer copes, to some extent, with the syntax of the source and the target languages.

At decoding time, since the input string to be translated consists of a sequence of words, we have to convert each source phrase of the transducer into a sequence of words. This conversion is straightforward in terms of a composition of the rough transducer with a left-to-right model (depicted in figure 5(b)). The integrated left-to-right transducer consists of a consecutive sequence of transitions analyzing word by word the source string, with the empty output and a probability equal to one in all the transitions except for the last one, where the output is the complete target phrase and the transition has a probability equal to that in the phrase-based model. Notice that the phrase-structure remains unchanged after the integration of the left-to-right word based model. In addition, on-the-fly integration is an efficient technique to reduce spatial costs. It might be noted that the conversion into words is not necessary when the input is speech instead of text, and this allows for an efficient use of memory.


q
$s_is_{i+1}...s_{i+I}|t_jt_{j+1}...t_{j+J}$
p
q'
q
$s_i|\lambda$
1
q1
$s_{i+1}|\lambda$
1
q2
q1
$s_{i+I}|t_jt_{j+1}...t_{j+J}$
p
q'

(a) Phrase-based SFST.

(b) On-the-fly integration of word-based model.

Figure 5: The integration of a left-to-right word based model in an edge of the phrase-based transducer. (a) An edge in the graph-diagram of the phrase-based finite-state transducer consists of the source phrase s_i^{i+I}, the target phrase t_j^{j+J} and it has associated a probability p. (b) The left-to-right word-based model keeps the general structure given by the phrase-based model, since neither the probability nor the input/output change. λ stands for the empty word.

As previously mentioned, the inference starts from a given segmented training corpus. This segmentation might be carried out taking the bilingual corpus into account or just from isolated monolingual parts, there is no restriction on this respect. Monolingual segmentation was tested in this work, nevertheless, other methods such as THOT toolkit [20] might be explored in order to obtain bilingual segmentations. Needless to say, segmentation plays an important role in this approach, thus, two possibilities are explored, linguistic and statistically motivated one in turn.

4.1. Linguistically motivated segmentation

Linguistic phrases were identified by Ametzagaña group following the next steps:

1. First of all, a morpho-semantic parsing allows to assign one or more tags to each word of the corpus. These tags include information about linguistic categories such as number, declension case, verb tense and aspect, etc. Besides, the stem and morphemes are identified. At this point, an ambiguous word would be assigned more than one tag-set.

2. A syntactic parsing allows to remove ambiguities under the following boundary: all words within a sentence have to share compatible categories. Regular expressions and regulated exceptions are also taken into account so as to select the appropriate sets of categories.

3. Once the syntactic and semantic parsing of each element is carried out unambiguously, linguistic phrases can be identified under a elementary criteria: group, recursively, all the words which share the same syntactic function whenever the frequency of that segment in the corpus exceeds a threshold. At first, just noun and verb phrases are distinguished, then, as the analysis goes ahead, more accurate groups such as composed stems, verbal periphrasis etc. are identified.

4.2. Statistically motivated segmentation

In order to obtain these segments, a simple procedure based on N-gram frequencies was used. This process is summarized in the following steps:

1. Given the training corpus, identify and extract all the 2-grams, 3-grams, . . ., *n*-grams available. In our case, we chose $n = 4$.

2. Sort them in order of decreasing values of *n* (n-grams before (n-1)-grams,. . . , 3-grams before 2-grams) and decreasing number of appearances.

3. For each sentence in the training set, get the subset of word *n*-grams that, while keeping the original order, satisfies a minimum number of occurrences. We chose 50 as a threshold. The idea is to replace all the

appearances of a sequence of words corresponding to an *n*-gram with a single unit obtained by joining all the words forming that *n*-gram. Some of the word *n*-grams might not appear after this process or might not satisfy the required minimum number of occurrences, due to the fact that they could be included in previous word *n*-grams with a higher value of *n*. The first of those *n*-grams not satisfying the required minimum number of occurrences is then removed. The process of relabeling and searching for not valid *n*-grams is iteratively repeated until getting a consistent segmentation.

5. Experimental results

5.1. Task and corpus

METEUS is a text and speech corpus in Basque and Spanish consisting of weather forecast reports picked up from the Internet [21] and later checked and expanded with: lemmas, statistical TAGS [16], and POS extracted with *FreeLing* toolkit [22] for Spanish, and for Basque courtesy of Ametzagaña group. The main features are shown in Table 1.

		Spanish	**Basque**
Training	Pair of sentences	14,615	
	Different pairs	8,445	
	Running words	191,156	187,195
	Vocabulary	702	1,135
	Singletons	162	302
	Average length	13.1	12.8
Test-1	Pair of sentences	1,500	
	Different pairs	1,173	
	Average length	12.6	12.4
	Perplexity (3-grams)	3.6	4.3
Test-2	Pair of sentences	1,800	
	Different pairs	500	
	Average length	17.4	16.5
	Perplexity (3-grams)	4.8	6.7

Table 1: Main features of METEUS corpus.

The corpus consists of a training set and two disjoint test sets. Test-1 keeps the statistics of the training set, and this means that it is representative of the task considered in the training set, while Test-2 is a training-independent set related to the same task (and thus, suitable as a benchmark in order to establish the lower threshold of the system). The latter was selected for speech translation evaluations. As Table 1 shows, Test-2 consists of 500 different sentences, being each one recorded for at least 3 speakers, we got as a result a total of 1,800 utterances by 36 speakers for each language.

5.2. Evaluation and confidence

The proposed translation models were assessed with the mentioned test sets under the typical automatic evaluation measures: BLEU, NIST, WER, PER. The goal of this work is to compare the performance of word, category and phrase-based SFSTs under either linguistic or statistical approaches. In order to be able to make fair comparisons between the different approaches the evaluation was carried out with 1,000 bootstrap test-sets. Table 2 shows the mean value and the 95% confidence interval for each of the aforementioned evaluation measures. To find out more about *confidence intervals* turn to [23, 24, 25].

Given the original test-set, D, consisting of N sentences, a *bootstrap test-set* (bootstrap sample) D*, is a set created by randomly selecting with replacement N sentences from D [26]. In D* there is nearly always duplication of individual sentences from D, in other words, it is likely that D* would include several sentences of D repeatedly while other sentences would be missed.

Experimentally, the mean value of the scores over a big-enough amount of bootstrap sets is close to the score obtained over the original set. Nevertheless, not all the bootstrap sets offer close scores to the mean (see Figure 6). Provided the sentences of the set are independent, the obtained bootstrap-scores will follow a Gaussian distribution. Following the notation on Table 2, the 95% of the bootstrap-scores are within $\mu \pm 2\sigma$ interval (being σ^2 the variance, and μ the mean value of the samples).

As shown in Table 2, for speech translation, both class-based and phrase-based models on their statistical approach offer quite close scores to the baseline. We might ask whether the reported differences are significant or not to draw a conclusion about their performance. That is, the arising question is whether we can assert with high confidence or not that class-based or phrase-based model is better than the baseline for speech translation. The Figure 6 shows the BLEU score of the 1,000 bootstrap test-sets for the baseline, the class-based and phrase-based systems. The graph shows that the results for the class-based approach and the baseline are quite close and we might want to know to what extent are them distinguishable. On the following we attempt at making out quantitatively this degree of uncertainty.

Whenever there is no overlapping between the 95% confidence intervals of two systems, then we can draw the conclusion with 95% certainty that for a given test, both systems would significantly differ. But when there is overlapping, instead of counting the amount of test-sets that gave as a result a score (as we did in Figure 6), we might count the number of times that one system outperforms the other for a given bootstrap set, and thereby, measure the *probability of improvement* (*poi*) [23, 24]. The *poi* is calculated by means of *paired-bootstrap*, which aims at measuring the discrepancy between two systems for a big amount of bootstrap-sets and then counting the number of times that such a difference represents that one system outperforms the other. For instance, let us assume that for the i-th bootstrap set the BLEU-score obtained for two systems under evaluation are $BLEU^{(i)}_{sys_1}$ and $BLEU^{(i)}_{sys_2}$ respectively. Intuitively, if the discrepancy

		Word		Category				Phrase			
				ling		stat		ling		stat	
		μ	2σ	μ	2σ	μ	2σ	μ	2σ	μ	2σ
Text Translation Test-1	BLEU	57.9	1.7	60.3	1.7	58.9	1.6	66.1	1.8	62.6	1.8
	NIST	7.4	0.1	7.6	0.1	7.5	0.1	8.1	0.1	7.8	0.2
	WER	32.8	1.5	31.4	1.6	32.3	1.7	27.6	1.7	29.9	1.6
	PER	27.7	1.3	26.6	1.3	27.1	1.5	22.3	1.4	24.3	1.3
Text Translation Test-2	BLEU	41.1	1.3	41.6	1.2	42.0	1.2	43.6	1.2	41.4	1.2
	NIST	6.0	0.1	6.0	0.1	6.1	0.2	6.3	0.1	6.0	0.1
	WER	47.5	1.2	48.0	1.2	47.5	1.1	48.0	1.3	51.0	1.4
	PER	39.4	1.1	40.4	1.0	39.4	1.1	38.9	1.1	41.1	1.2
Speech Translation Test-2	BLEU	38.5	1.2	38.9	1.2	38.8	1.2	40.2	1.2	40.0	1.4
	NIST	5.7	0.1	5.8	0.1	5.7	0.1	5.9	0.1	5.9	0.1
	WER	51.3	1.3	50.5	1.3	51.4	1.3	50.5	1.3	53.9	1.4
	PER	42.5	1.10	41.8	1.0	42.4	1.1	41.1	1.1	44.1	1.3
Recognition	WER	8.3	0.4	7.3	0.4	8.2	0.3	9.6	0.5	12.1	0.7

Table 2: Text-to-text and speech-to-speech translation results with different SFST models, namely, word-based, category-based and phrase-based model. For the latter two models both linguistic (ling) and statistical (stat) approaches are explored. The mean value for 1,000 bootstrap sets (μ), and the 95% confidence interval $[\mu - 2\sigma, \mu + 2\sigma]$.

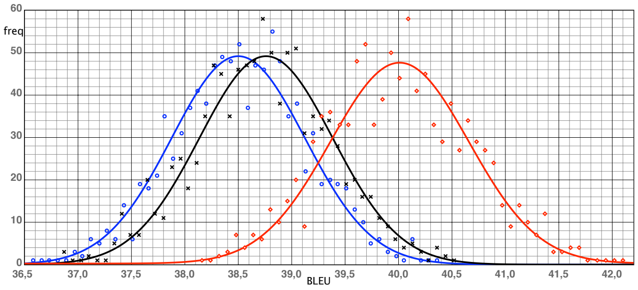


Figure 6: The ordinate axis represents the amount of bootstrap-sets (out of 1,000) that gave as a result the BLEU score in the abscissa. Speech translation results with the baseline system (experimental values drawn with circles), statistical approach of class-based system (experimental values drawn with crosses) and statistical approach for phrase-based system (experimental values drawn with squares). The curves are just the Gaussian functions that better fit the samples.

between the two systems is positive, meaning expression (6), then sys_1 system outperforms sys_2 on the i-th bootstrap test-set.

$$\Delta BLEU^{(i)}_{(sys_1,sys_2)} = BLEU^{(i)}_{sys_1} - BLEU^{(i)}_{sys_2} \qquad (6)$$

Let us measure the discrepancy in a score of the two systems over a big amount of bootstrap test-sets (that is, a big B). If sys_1 is considered to be better than sys_2 for b bootstrap sets out of B, the *poi* of sys_1 with regard to sys_2 can be approached by $poi \simeq \frac{b}{B}$ as eq. (7) suggests in more general terms.

$$poi(\Delta Score_{(sys_1,sys_2)}) \simeq \frac{1}{B}\sum_{i=1}^{B}\Theta(Score^{(i)}_{sys_1} - Score^{(i)}_{sys_2}) \qquad (7)$$

Where $Score^{(i)}_{sys_1}$ is the score obtained with the system sys_1 over the i-th bootstrap set; and whenever the *Score* is an accuracy value (such as BLEU or NIST) where the higher the better performance, then $\Theta(x)$ is the *Heaviside unit step function*, denoted as $H(x)$ in expression (8), otherwise, when the *Score* is an error rate (such as WER or PER) where the higher the worse, then $\Theta(x) = H(-x)$.

$$H(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases} \qquad (8)$$

Since the major uncertainty over the discrepancies between the mentioned systems are related to speech translation of Test-2, we resorted to *poi* in that case. Table 3 shows the *poi* of class-based (CB) and phrase-based (PB) models with regard to the baseline (WB), as well as the *poi* for the linguistic approach with respect to the statistical one under either CB or PB approaches.

5.3. Discussion

Having a look at Tables 2 and 3, we conclude that there is not a complete agreement between all the automatic evaluation scores. For instance, according to the BLEU, the statistical approach of class based model seems to perform better for speech translation than the baseline, however, it is the other way around according to the WER. Let us mention that according to the ACL-2007 evaluation campaign on machine translation, BLEU score has, by far, better correlation with human judgments than WER.

poi		**Score**			
$\mathbf{sys_1}$	$\mathbf{sys_2}$	BLEU	NIST	WER	PER
CB-stat	WB	0.812	0.641	0.385	0.461
CB-ling	WB	0.844	0.955	0.982	0.958
CB-ling	CB-stat	0.667	0.936	0.995	0.973
PB-stat	WB	0.996	0.999	0.000	0.002
PB-ling	WB	0.999	1.000	0.934	0.997
PB-ling	PB-stat	0.612	0.527	1.000	1.000

Table 3: Probability of improvement of sys_1 respect to sys_2 on the basis of several Scores (namely, BLEU, NIST, WER and PER), that is, $poi(\Delta Score_{(sys_1,sys_2)})$. The systems are denoted as follows: WB stands for word-based model, CB-stat and CB-ling stand for statistical and linguistic approach of class-based model respectively, and analogously, PB-stat and PB-ling stand for statistical and linguistic approach of phrase-based model respectively. The results correspond to speech translation experiments of 1,000 bootstrap-sets of Test-2.

The results in Table 2 show that phrase-based SFSTs (in either statistical or linguistic approach) outperform word-based SFSTs. Following the automatic evaluation class-based approach is just slightly better than the baseline, being the differences not statistically significant for the majority of the situations studied (text-to-text translation of Test-1 or Test-2 and speech-to-speech translation of Test-2). Nevertheless, a manual inspection over 50 randomly extracted sentences turned out that class-based approach outperformed the baseline regarding the meaning transfer but there was almost no difference in what fluency concerned. Regarding the linguistic and statistical approaches, the linguistic one has resulted in slightly better translation results than the statistical one for both class-based and phrase-based models.

With respect to the speech translation results, note that each translation model is accompanied with a recognition score (see Table 2). This is due to the fact that speech translation was carried out using the so-called *integrated architecture* [27], which involves a tight on-the-fly integration between acoustic and translation model. Both the recognized string in the source language and its translation are jointly obtained in a single decoding. As a matter of comparison, the speech recognition score for this task using a 3-TSS language model is WER = 7.9. Therefore, some translation systems not only improve the translation scores (respect to the baseline) but also the recognition scores respect to a classical speech recognition system.

6. Concluding remarks and future work

In this work we have made an overview of one approach of the classical finite-state transducers, namely GIATI. That approach has been enhanced in two different ways, by the so-called category-based and phrase-based approaches respectively. In addition, both approaches have been studied taking both linguistic and statistical information into account. A thorough study with confidence measures has been carried out. As a result, we conclude that both class and phrase-based models outperform the baseline for this task, and thus it will be worth investigating the proposed general methods on more complex tasks. With regard to the statistical and linguistic knowledge sources, in this work linguistic ones got better results but it might be of interest to study the combination of both of them.

For future work we aim at exploring other kind of statistical phrases. On the other hand, category-based approach with different number of statistical categories might also be considered. Furthermore, we will attempt at carrying out a tight combination of both categorization and phrase-based approaches. In essence, our goal is to make use of categorization over phrases instead of running words as suggested in [28] for speech recognition, applied, in this case, to speech translation.

7. Acknowledgments

We would like to thank the Ametzagaiña group (`http://www.ametza.com`) and *Josu Landa* in particular, for providing us with the morpho-syntactic parse that made possible this work.

This work has been partially supported by the University of the Basque Country under grant 9/UPV00224.310-15900/2004 and by CYCIT under grant TIN2005-08660-C04-03

8. References

[1] E. Vidal, F. Thollard, F. C. C. de la Higuera, and R. Carrasco, "Probabilistic finite-state machines - part II," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 7, pp. 1025–1039, 2005.

[2] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, June 1990.

[3] F. Casacuberta and E. Vidal, "Learning finite-state models for machine translation," *Machine Learning*, vol. 66, no. 1, pp. 69–91, 2007.

[4] M. Mohri, "Finite-state transducers in language and speech processing," *Computational Linguistics*, vol. 23, no. 2, pp. 269–311, 1997.

[5] D. Caseiro and I. Trancoso, "A specialized on-the-fly algorithm for lexicon and language model composition." *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1281–1291, 2006.

[6] F. C. N. P. Mehryar Mohri and M. D. Riley, "AT&T FSM LibraryTM Finite-State Machine Library," 2003, `www.research.att.com/sw/tools/fsm`

[7] F. Casacuberta and E. Vidal, “Machine translation with inferred stochastic finite-state transducers,” *Computational Linguistics*, vol. 30, no. 2, pp. 205–225, 2004.

[8] P. García and E. Vidal, “Inference of k-testable languages in the strict sense and application to syntactic pattern recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 9, pp. 920–925, 1990.

[9] M. I. Torres and A. Varona, “k-TSS language models in speech recognition systems,” *Computer Speech and Language*, vol. 15, no. 2, pp. 127–149, 2001.

[10] F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, Mar. 2003.

[11] P. F. Brown, V. J. Della Pietra, P. V. de Souza, J. C. Lai, and R. L. Mercer, “Class-based n-gram models of natural language,” *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[12] T. Niesler, E. Whittaker, and P. Woodland, “Comparison of part-of-speech and automatically derived category-based language models for speech recognition,” in *ICASSP’98, Seattle.*, 1998, pp. 177–180.

[13] J. Amengual, J. Benedí F. Casacuberta, A. Castaño, A. Castellanos, V. Jimenez, D. Llorens, A. Marzal, F. Prat, E. Vidal, and J. Vilar, “Using categories in the eutrans system,” in *ACL-ELSNET Workshoop on Spoken Language Translation*, Madrid (Spain), july 1997, pp. 44–53.

[14] J. Amengual, J. Benedí, F. Casacuberta, M. Castao, A. Castellanos, V. Jiménez, D. Llorens, A. Marzal, M. Pastor, F. Prat, E. Vidal, and J. Vilar, “The EuTrans-I speech translation system,” *Machine Translation*, vol. 1, 2000.

[15] A. Pérez, M. I. Torres, and F. Casacuberta, “Towards the improvement of statistical translation models using linguistic features,” in *Proceedings of the FinTAL. Lecture Notes in Computer Science 4139*, August 2006, pp. 716–725.

[16] F. J. Och, “An efficient method for determining bilingual word classes,” in *EACL ’99: Ninth Conf. of the Europ. Chapter of the Association for Computational Linguistics*, Bergen, Norway, June 1999, pp. 71–76.

[17] A. Pérez, M. I. Torres, and F. Casacuberta, “Speech translation with phrase based stochastic finite-state transducers,” in *ICASSP 2007*. IEEE, April 15-20 2007.

[18] S. Kumar, Y. Deng, and W. Byrne, “A weighted finite state transducer translation template model for statistical machine translation,” *Natural Language Engineering*, vol. 12, pp. 35–75, December 2005.

[19] B. Zhou, S. Chen, and Y. Gao, “Constrained Phrase-based Translation Using Weighted Finite State Transducer,” in *ICASSP 2005*, vol. 1, 2005, pp. 1017–1020.

[20] D. Ortiz, I. Garca-Varea, and F. Casacuberta, “Thot: a toolkit to train phrase-based statistical translation models,” in *X MT Summit*. Phuket, Thailand: Asia-Pacific Association for Machine Translation, September 2005, pp. 141–148.

[21] A. Pérez, M. I. Torres, F. Casacuberta, and V. Guijarrubia, “A Spanish-Basque weather forecast corpus for probabilistic speech translation,” in *Proceedings of the 5th Workshop on Speech and Language Technology for Minority Languages (SALTMIL)*, Genoa, Italy, 2006.

[22] X. Carreras, I. Chao, L. Padró, and M. Padró, “Freeling: An open-source suite of language analyzers,” in *In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC’04)*, Lisbon, Portugal, 2004, `http://garraf.epsevg.upc.es/freeling`.

[23] M. Bisani and H. Ney, “Bootstrap estimates for confidence intervals in ASR performance evaluation,” in *ICASSP 2004*, vol. 1, 2004, pp. 409–412.

[24] Y. Zhang and S. Vogel, “Measuring confidence intervals for the machine translation evaluation metrics,” in *Proceedings of The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, October 2004.

[25] P. Koehn, “Statistical significance tests for machine translation evaluation,” in *Proceedings of EMNLP 2004*, D. Lin and D. Wu, Eds. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 388–395.

[26] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY: John Wiley and Sons, 2000.

[27] F. Casacuberta, E. Vidal, and J. M. Vilar, “Architectures for speech-to-speech translation using finite-state models,” Philadelphia, USA, 2002, pp. 39–44.

[28] R. Justo and M. I. Torres, “Phrases in category-based language models for Spanish and Basque ASR,” in *Proceedings of the Interspeech07*, Antwerp, Belgium, August 27-31 2007, pp. 2377–2380.