Evaluation of Alternatives on Speech to Sign Language Translation

R. San-Segundo[1], A. Pérez[2], D. Ortiz[3], L. F. D'Haro[1], M. I. Torres[2], F. Casacuberta[3]

[1]Grupo de Tecnología del Habla. Universidad Politécnica de Madrid. Spain.
[2]Dpto de Electricidad y Electrónica. Facultad de Ciencia y Tecnología. Universidad del País Vasco. Spain.
[3]Instituto Tecnológico de Informática. Universidad Politécnica de Valencia. Spain.
lapiz@die.upm.es

Abstract

This paper evaluates different approaches on speech to sign language machine translation. The framework of the application focuses on assisting deaf people to apply for the passport or related information. In this context, the main aim is to automatically translate the spontaneous speech, uttered by an officer, into Spanish Sign Language (SSL).

In order to get the best translation quality, three alternative techniques have been evaluated: a rule-based approach, a phrase-based statistical approach, and a approach that makes use of stochastic finite state transducers. The best speech translation experiments have reported a 32.0% SER (Sign Error Rate) and a 7.1 BLEU (BiLingual Evaluation Understudy) including speech recognition errors.

Index Terms: Machine Translation, Spanish Sign Language, Speech Translation.

1. Introduction

Spoken language translation is being investigated in a number of join projects like C-Star, ATR. Vermobil, Eutrans, LC-Star, PF-Star and TC-Star. The best performing translation systems are based on various types of statistical approaches [1], including example-based methods [2], finite-state transducers [3] and other data driven approaches. In restricted domains, the rule-based approaches have demonstrated to work very well with a low rule-development effort [4].

In the recent years, several groups have showed interest in machine translation into Sign Languages, developing several prototypes: example-based [5], rule-based [6], full sentence [7] or statistical [8] approaches. This paper includes the first experiments on one of the precursors speech to Sign Language animation translation systems, and indeed the first one developed specifically for the Spanish Sign Language (SSL).

Even though speech input machine translation can be implemented in a single translation model that integrates the acoustic models within the translation model [3], in this work the problem is tackled by means of the typical speech decoder in a serial architecture with a text-to-text translation model in order to evaluate all the alternatives in the same conditions. This paper focuses on the translation step where three different alternatives have been evaluated: a rule-based model, and two different data-driven models (a statistical phrase-based model and a statistical model based on Stochastic Finite State Transducers (SFSTs)). The strength of each of the three translation approaches under corrupted inputs has been studied. In general, text-to-text translation systems are built on the supposition that the input sentence will be correct. However, in a speech translation system the translation model has to deal with the output of the speech recognition system, which may be not exempt from errors.

The motivation of this work arose from the Albayzín Evaluation Campaign organized in November 2006 by the Spanish National Network on Speech Technology. The underlying work has also faced previously unexplored problems in this field such as new proposals to deal with out of vocabulary words (OOVs). In fact, another contribution of this paper lays on the analysis of the capability of state of the art methods in machine translation to deal with scarce resources such as tasks involving Sign Language.

The remaining of this paper is organized as follows: section 2 describes the task and the available data; section 3 summarizes the automatic speech recognition system used to decode input speech; sections 4, 5 and 6 are devoted to describe a rule-based model, a phrase-based model and a finite-state model respectively, the three different approaches taken into account; the experimental results are summarized in section 7; some conclusions are outlined in section 8.

2. Domain and database

The experiments under study are focused on speech to sign language translation in a limited domain. The experimental framework is restricted to the sentences spoken by an officer when assisting people in applying for the National Identification Document or related information. In this context, a speech to sign language translation system is very useful since most of the officers do not know sign language and they have difficulties when interacting with deaf people. This system translates the officer explanations into sign language animation to aim a better service at deaf people.

For developing purposes, a specific corpus has been collected. The most used sentences have been selected from typical dialogues between officers and users, adding up a total of 416 sentences that contain more than 650 different words.

In order to represent the Spanish Sign Language (SSL) in text symbols, a suitable encoding has been developed. Each sign has been represented by a word written in capital letters: e.g. "*you have to pay 20 euros as document fee*" is translated into "*FUTURE YOU PAY TWENTY EURO DOC_FEE*""

Table 1. Main features of the bilingual corpus in Spanish and Spanish Sign Language (SSL).

		Spanish	**SSL**
Training	Sentence pairs	266	
	Different sentences	259	253
	Running words	3,153	2,952
	Vocabulary	532	290
Test	Sentence pairs	150	
	Running words	1,776	1,688
	Unknown words, OOV	93	30
	Perplexity (3-grams)	15.4	10.7

Once the SSL encoding was established, an expert translated the original set into SSL, making use of more than 300 different signs. Then, the 416 pairs were randomly divided in two disjoint sets: 266 for training and 150 for testing purposes. The main features of the corpus are summarized in Table 1. For both text-to-sign and speech-to-sign translation purposes the same test set has been used. As the application has been conceived to be realistic, the speech recognizer must be speaker independent. Thus, 14 speakers were recorded (7 male and 7 female). Each test sentence was pronounced by at list 4 speakers, obtaining, as a result, a total of 700 utterances.

As it is shown in Table 1, the size of the vocabulary in comparison with the overall amount of running words in the training set is very high (17%). In addition, the perplexity of the test set is high considering the small size of the vocabulary. The mentioned ratio and the high perplexity are unquestionable signs of data scarcity, which is likely to cause a kind of dispersion when estimating the parameters of the statistical translation models.

In these circumstances, the high amount of unknown words in the test set (OOVs) represents another important issue. In this task, there are 93 OOVs out of 532. A usual statistical system is not provided with morpho-syntactic parsers to analyze the unknown words, therefore, they can hardly manage with them. So far, in the literature only naive methods have been implemented to face the translation of OOVs. The commonly adopted solution displays the input unknown word itself, without any change, in the output language. This heuristic is successful on the assumption that the most of the unknown words are proper names, numbers, etc. That is, OOVs are in principal any sort of token that can be transcribed in the same way in any language. In the present task the OOVs are Spanish running words (no proper words), which indeed do not match any symbol on Sign Language. Therefore, the common solution was of no help under this framework, and thus, new solutions had to be proposed

3. Speech Recognition Results

The speech recognizer is a state of the art recognizer developed at GTH-UPM [9]. It is a HMMs-based (Hidden Markov Models) system that recognizes continuous speech from any Spanish speaker. It also generates a confidence measure for each recognized word [10].

With regard to the *acoustic modeling*, the ASR uses 5760 triphone HMMs for modeling context dependent allophones and 16 silence and noise HMMs for detecting acoustic effects (non speech events like background noise, speaker artifacts, filling pauses, etc). The *language model* is just a bigram language model due to the data sparseness. There are a few sentences to train the model compared to the size of the vocabulary, and with the amount of different n-grams.

Table 2. Speech recognition results.

WER	Ins (%)	Del (%)	Sub (%)
24.08	2.61	6.71	14.76

The results of the Table 2 show the outstanding influence of data sparseness (due to the small amount of data) over the decoding process. As a lower threshold, considering an ideal ASR which would recognize accurately every single known word, there would be several errors due to OOVs: WER=5.23, Ins=0, Del=0 and Sub=5.23. Provided that the system has no ability to generate other words on the target language than those seen in the training set, the presence of OOVs on the reference set implies that the lowest WER for this system is strictly greater than 0.

Next sections describe the three translation alternatives.

4. Rule-based Translation

In this approach, the natural language translation module has been implemented using a rule-based technique considering a bottom-up strategy. In this case, the relations between signs and words are defined by hand. In a bottom-up strategy, the translation analysis is performed starting from each word individually and extending the analysis to neighborhood context words or already-formed signs (generally named blocks). This extension is done to find specific combinations of words and/or signs (blocks) that generate another sign. The rules implemented by the expert define these relations.

The translation process is carried out in two steps. In the first one, every word is mapped into one or several syntactic-pragmatic tags. After that, the translation module works applying different rules that convert the tagged words into signs by means of grouping words or signs (blocks) and defining new signs. At the end of the process, the block sequence must correspond to the sign sequence resulting from the translation process (Figure 1). Considering the four situations reported in [11], it is possible to classify the rules in four types: one word corresponds to an specific sign, several words generate a unique sign, one word generates several signs, and the last kind of rules are those that generate several signs from several words with certain relationships between them The final version of the rule base translation module contains 170 translation rules.

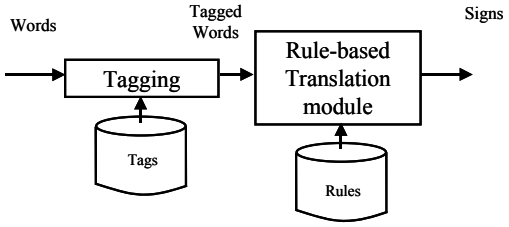

Figure 1. Rule-based Translation process

4.1. Sign confidence measure

The translation module generates one confidence value for every sign. This sign confidence is computed by an internal procedure that is coded inside the proprietary language interpreter that executes each rule. In this internal engine, there are primitive functions, responsible for the execution of the rules written by the experts. Each primitive has its own way to generate the confidence measure for the elements it produces (computed from the confidence measures of the primitive inputs).

5. Phrase-based Translation

The Phrase-based translation system is based on the software released to support the shared task at the 2006 NAACL Workshop on Statistical Machine Translation (http://www.statmr.org/wmt06/). The phrase model has been trained following these steps:

1. **Word alignment computation.** At this step, the GIZA++ software [12] has been used to calculate the alignments between words and signs (considering the training set).

The parameter "*alignment*" was fixed to "*grow-diag-final*" as the best option.

2. **Phrase extraction** [13]. All phrase pairs that are consistent with the word alignment are collected. The maximum size of a phrase has been fixed to 7. The number of phrases considered (with length between 1 a 7) has been 3433.
3. **Phrase scoring**. The translation probabilities are computed for all phrase pairs. Both translation probabilities are calculated: forward and backward.

The Pharaoh decoder is used for the translation process. This program is a beam search decoder for phrase-based statistical machine translation models [14]. In order to obtain a 3-gram language model needed by Pharaoh, the SRI language modeling toolkit has been used. The Carmel software were used for n-best list generation.

6. Stochastic Finite State Transducers

Stochastic finite state transducers (SFST) have proved to be useful in language processing and in automatic speech recognition systems. They have also been proposed for statistical machine translation applications. An *stochastic* finite-*state transducer* is a tuple $T = (\Sigma, \Delta, Q, q_0, R, F, P)$ where:

Σ is a finite set of input symbols (source words);
Δ is a finite set of output symbols (target words);
Q is a finite set of states;
$q_0 \in Q$ is the initial state;
$R \subseteq Q \times \Sigma \times \Delta^* \times Q$ is a set of transitions such as $(q, s, \tilde{t}, q')$, which is a transition from the state q to the state q', with the source word (or phrase [16]) s and producing the substring $\tilde{t}$;
$P : R \rightarrow [0, 1]$ probability distribution over transitions;
$F : Q \rightarrow [0, 1]$ final state probability distribution;

The probability distributions satisfy the constraint:

$$\forall q \in Q \quad F(q) + \sum_{\forall s, \tilde{t}, q'} P(q, s, \tilde{t}, q') = 1 \qquad (1)$$

The SFST is characterized by both the topology and the probability distributions. These distinctive features can be automatically learnt from bilingual corpora by efficient algorithms, such as GIATI (Grammar Inference and Alignments for Transducers Inference) reported at [3]. In this work a k-TSS topology [15] has been selected which provides the SFST with a syntactic back-off smoothing. The back-off allows the SFST to deal with events (n-grams) that have not been seen amongst the training data.

$$\hat{t} = \arg\max_{t} P(\mathbf{s}, \mathbf{t}) \approx \arg\max_{t} \max_{d(\mathbf{s},\mathbf{t})} P(d(\mathbf{s}, \mathbf{t})) \quad (2)$$

With the transducer and an input sentence $s \in \Sigma^+$, the translation process implies the searching for the most likely output string $\hat{t} \in \Delta^*$ through all the possible output strings, as equation (2) summarizes. Where $d(\mathbf{s},\mathbf{t})$, represents a path in the SFST, compatible both with the input sentence **s** and the output **t**. Therefore, the searching criteria in the SFST deals with the joint probability of sentence pairs.

6.1. Syntactic-based approach to deal with OOVs

So far, a mechanism to deal with unknown n-grams has been implemented: the back-off smoothing. However, an additional framework that enables the analysis of out of vocabulary words has to be included (being an OOV, every word in the input language that does not belong to the input vocabulary).

One possible way to interpret unknown words is by means of syntax, that is, taking into account the context where that word appears. Both left and right contexts entail information about the unknown word. Even if a word has never been seen before, the meaning might be guessed thanks to the context, and even a synonym of the unknown word could be proposed.

Finite-state models offer an appropriate framework to emulate this behavior. At decoding time, when an unknown word appears, all the known words are explored as possible synonyms of the unknown word. At that point, many alternative paths have to be explored, but little by little, as the analysis goes ahead, the probability associated to most of the paths decrease drastically, eventually only a few seem to be suitable. Nevertheless, those paths that seem futile could be pruned to get a faster decoder.

Alternatively, it is possible to attach a probability that states how likely is a word of the vocabulary, $v_i \in \Sigma$, to be the synonym of the given unknown word: $P_d(v_i|OOV)$ where d stands for a distance measure, such as, the Levenshtein distance (also referred to as edit distance). Under Levenshtein distance approach, words that share the same stem would take priority over other words of the vocabulary. Nevertheless, for morphologically rich languages may be a useless measure.

A suitable metric for the particular task and language under consideration should be explored on the basis of the knowledge provided by either dictionary or POS (Part Of Speech) information or statistical tags. In this work, a constant distance measure has been selected. This distance assumes all the words of the vocabulary have the same probability to be the translation of the unknown word. In further work, the influence on the performance of different distance measures will be explored.

7. Translation results

The evaluation has been carried out with the test set presented in Table 1. In order to assess the quality of the obtained translations, four evaluation measures have been taken into account: SER (Sign Error Rate), PER (Position Independent SER), BLEU (BiLingual Evaluation Understudy), and NIST. Let us notice that the former two measures are error measures (the higher value, the worse quality) whereas the latter two are accuracy measures (the higher, the better). As a baseline (denoted in the Table 3 as Text), the text-to-text translation results are included, by means of considering directly the transcription of the utterance.

Table 3. Translation results.

Text	**SER**	**PER**	**BLEU**	**NIST**
Rule-based	16.75	13.17	0.7217	8.5992
SFST	29.21	25.48	0.5801	7.4042
Phrase-based	33.74	29.14	0.5152	6.6505
Speech	**SER**	**PER**	**BLEU**	**NIST**
Rule-based	**31.99**	**27.44**	**0.5553**	**7.0862**
SFST	38.47	33.82	0.5139	6.7108
Phrase-based	39.02	34.45	0.4831	6.2143

The SER is higher when using the speech recognition output instead of the transcribed sentence. The reason is the

speech recognizer introduces recognition errors that produce more translation errors: the percentage of wrong signs increases and, consequently, the BLEU decreases.

Comparing the performance of the systems when the input is a well formed sentence (Text) with that obtained when the input sentence has been corrupted due to a wrong speech recognition, the rule-based system is the most sensitive one, being its BLEU decreasing of 0.1664, whereas for the other statistical systems the decreasing is 0.0662 (SFST) and 0.0321 (PB) respectively. However, the rule-based approach is by far the most successful one for this task.

Analyzing the results, the most frequent errors committed by the translation module have the following causes:

- Unknown words. There is a high number of OOVs in the test set to deal with.
- Omission of the subject in SSL. In Spanish, it is very common to omit the subject of a sentence, but in SSL it is mandatory.
- Several possible translations. One sentence can be translated into different sign sequences. The system is unfairly penalized in several examples where the passive form is omitted. Therefore, multiple references would be desirable in order to offer more accurate results.
- Ambiguities. In Sign Language, a verb complement is represented by a specific sign: for example, a time complement is introduced with the sign WHEN, and a mode complement is introduced with the sign HOW. Sometimes is very difficult to classify the type due to ambiguities.
- Short distance word reordering. In Spanish, the plural number case is usually attached at the end of the stem, whereas in SSL, an additional sign is used, which, besides, precedes the sign to be modified.

8. Conclusions and future work

This paper presents an evaluation of different approaches for translating speech into sign language: a rule-based approach, a statistical approach based on a phrase-model, and a connectionist one using stochastic finite state transducers.

The error propagation through the speech translation system has also been evaluated, that is, from the speech recognizer in the first stage to the text translator in the second stage. The robustness of each translation model to overcome the errors due to the speech decoding has been studied. The rule-based strategy has shown the best results on this task. However, the development of the rules is difficult to extend to more general domains. Amongst the statistical approaches, the stochastic finite state transducers offer slightly better results with a very low model developing effort.

This analysis has been the result of the Albayzín Evaluation organized in Nov. 2006 by the Spanish National Network on Speech Technology. The results have reported a SER lower than 40% and a BLEU higher than 0.48 in all cases.

As data sparseness has proved to entail a real challenge when dealing with example-based methods, the main goal for further work is to quantitatively study the influence of OOVs in the translation performance. In order to fix this kind of errors, the methods sketched in this paper will be more thoroughly developed and applied in several tasks.

9. Acknowledgements

This work has been partially supported by the Spanish Network on Speech Technology, MEC under grant TEC2005-24712-E by the University of the Basque Country under grant 9/UPV00224.310- 15900/2004 and by CYCIT under grant TIN2005-08660-C04-03

10. References

[1] Och J., H. Ney. (2002). “Discriminative Training and Maximum Entropy Models for Statistical Machine Translation”. Annual Meeting of the Ass. ACL, Philadelphia, PA, pp. 295-302. 2002.

[2] Sumita E., Y. Akiba, T. Doi et al. (2003). “A Corpus-Centered Approach to Spoken Language Translation”. Conf. Of Ass. for Computational Linguistics (ACL) Hungary. pp171-174.

[3] Casacuberta F., E. Vidal. (2004). “Machine Translation with Inferred Stochastic Finite-State Transducers”. Comp. Linguistics, V30, n2, 205-225.

[4] M. Huenerfauth. 2004. A multi-path architecture for machine translation of English text into American Sign language animation. HLT-NAACL, Boston, MA, USA.

[5] S. Morrissey and A. Way. 2005. An example-based approach to translating sign language. In Workshop Example-Based Machine Translation (MT X–05), pages109–116, Phuket, Thailand, September.

[6] R. San-Segundo, R. Barra, L.F. D’Haro, J.M. Montero, R. Córdoba, J. Ferreiros. “A Spanish Speech to Sign Language Translation System”. Interspeech 2006.

[7] S. J. Cox, M. Lincoln, J Tryggvason, M Nakisa, M. Wells, Mand Tutt, and S Abbott. TESSA, a system to aid communication with deaf people. In ASSETS 2002, pages 205-212, Edinburgh, Scotland, 2002.

[8] J. Bungeroth and H. Ney: Statistical Sign Language Translation. In Workshop on Representation and Processing of Sign Languages, LREC 2004, 105-108.

[9] http://lorien.die.upm.es

[10] Ferreiros, J., R. San-Segundo, F. Fernández, L.F.D'Haro, V. Sama, R. Barra, P. Mellén. New Word-Level and Sentence-Level Confidence Scoring Using Graph Theory Calculus and its Evaluation on Speech Understanding. Interspeech 2005.

[11] San-Segundo R., J.M. Montero, J. Macías-Guarasa, R. Córdoba, J. Ferreiros, J.M. Pardo. (2004). “Generating Gestures from Speech”. ICSLP 2004.

[12] Franz Josef Och, Hermann Ney. "Improved Statistical Alignment Models". Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, Hongkong, China, October 2000.

[13] P. Koehn, F. J. Och, and Daniel Marcu, “Statistical Phrase-Based Translation” HLT/NAACL 2003.

[14] P. Koehn “Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models”, AMTA-04.

[15] M. Inés Torres and Amparo Varona, “k-tss language models in speech recognition systems,” Computer Speech & Language, vol. 15, no. 2, pp. 127–149, 2001.

[16] Alicia Pérez, M. Inés Torres, and Francisco Casacuberta, “Speech translation with phrase based stochastic finite-state transducers,” in Proceedings of the 32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), Honolulu, Hawaii USA, April 15-20 2007, IEEE.