AUTOMATIC CATEGORIZATION FOR IMPROVING SPANISH INTO SPANISH SIGN LANGUAGE MACHINE TRANSLATION

ABSTRACT

This paper describes a preprocessing module for improving the performance of a Spanish into Spanish Sign Language (Lengua de Signos Española: LSE) translation system when dealing with sparse training data. This preprocessing module replaces Spanish words with associated tags. The list with Spanish words (vocabulary) and associated tags used by this module is computed automatically considering those signs that show the highest probability of being the translation of every Spanish word. This automatic tag extraction has been compared to a manual strategy achieving almost the same improvement. In this analysis, several alternatives for dealing with non-relevant words have been studied. Non-relevant words are Spanish words not assigned to any sign. The preprocessing module has been incorporated into two well-known statistical translation architectures: a phrase-based system and a Statistical Finite State Transducer (SFST). This system has been developed for a specific application domain: the renewal of Identity Documents and Driver's License. In order to evaluate the system a parallel corpus made up of 4,080 Spanish sentences and their LSE translation has been used. The evaluation results revealed a significant performance improvement when including this preprocessing module. In the phrase-based system, the proposed module has given rise to an increase in BLEU (Bilingual Evaluation Understudy) from 73.8% to 81.0% and an increase in the human evaluation score from 0.64 to 0.83. In the case of SFST, BLEU increased from 70.6% to 78.4% and the human evaluation score from 0.65 to 0.82.

1. Introduction

In the world, there are around 70 million people with hearing deficiencies (information from World Federation of the Deaf: http://www.wfdeaf.org/). Deafness brings about significant communication problems: most deaf people have serious problems when expressing themselves in these languages or understanding written texts. They have problems with verb tenses, concordances of gender and number, etc., and they have difficulties when creating a mental image of abstract concepts. According to information from INE (Spanish Statistic Institute: http://www.ine.es), in Spain, there are 1,064,000 deaf people and 50% of them are more than 65 years old. 47% of the deaf population do not have basic studies or are illiterate, and only between 1% and 3% have finished their university studies, as opposed to 21% of Spanish hearing people (information from MEC: Spanish Ministry of Education: http://www.mec.es). Also, 20% of the deaf population is unemployed (30% for women). This fact can cause deaf people to have problems when accessing information, education, job, social relationship, culture, etc. It is necessary to make a difference between "deaf" and "Deaf": the first one refers to non-hearing people, and the second one refers to non-hearing people who use a sign language to communicate between themselves (their mother tongue), making them part of the "Deaf community". In Spain, around 35% of the deaf population uses a sign language for communicating between themselves. This population is the focus of this work. This percentage is relatively low because there are deaf people that became deaf after language acquisition (this percentage is higher for deaf people over 65) so they know written Spanish very well and they prefer it to sign language.

Sign languages are fully-fledged languages that have a grammar and lexicon just like any spoken language, contrary to what most people think. Traditionally, deafness has been associated to people with learning problems but this is not true. The use of sign languages defines the Deaf as a linguistic minority, with learning skills, cultural and group rights similar to other minority language communities. An important problem is that there are not enough sign-language interpreters. In the USA, there are 650,000 Deaf people (who use a sign language), although there are more people with hearing deficiencies, but only 7,000 sign-language interpreters, i.e. a ratio of 93 deaf people to 1 interpreter. Finland has the best ratio, 6 to 1, and Slovakia the worst with 3,000 users to 1 interpreter (Wheatley and Pabsch, 2010). In Spain this ratio is 221 to 1. This information shows the need to develop automatic translation systems with new technologies for helping hearing and Deaf people to communicate between themselves.

This paper describes a preprocessing module for improving a statistical translation system that helps Deaf people to communicate with government employees in a restricted domain: the renewal of Identity Documents and Driver's License. The proposed approach has demonstrated a very good performance for translating written Spanish into LSE (Lengua de Signos Española) thus reducing the translation error considerably. This new technique allows the system to be adapted better to the differences between written and sign languages. In section 3, these differences will be presented. This preprocessing module has been integrated into a Spoken Spanish to Spanish Sign Language system (San-Segundo et al., 2008). The system is made up of a speech recognizer (for decoding the spoken utterance into a word sequence), a natural language translator (for converting a word sequence into a sequence of signs belonging to the sign language), and a 3D avatar animation module (for playing back the signs) (Figure 1). This system has been designed to translate the government employee's explanations into LSE when government employees provide face-to-face services. This paper proposes to include a fourth module called "preprocessing" between the speech recognition and language translation modules (Figure 1). This preprocessing module replaces Spanish words with associated tags.

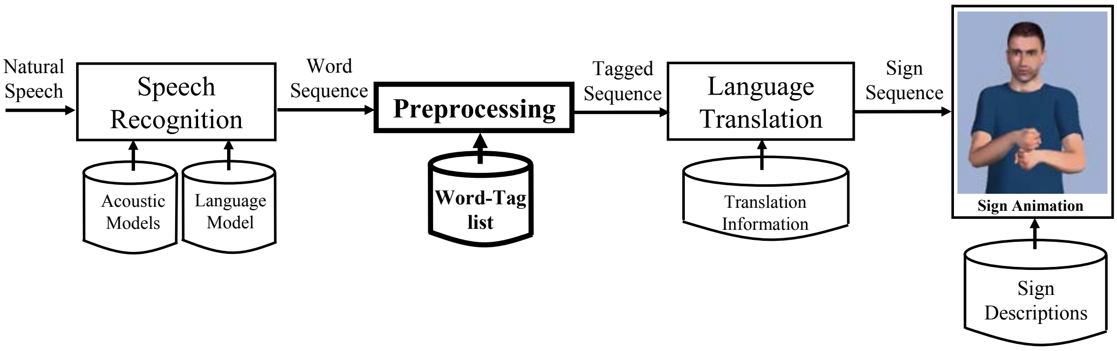

Figure 1. Spanish into LSE translation system

For the natural language translation module, two different statistical strategies have been analyzed: a phrase-based system (Moses) and a Statistical Finite State Transducer (SFST). The proposed preprocessing module has been incorporated into and evaluated with both translation strategies.

The paper is organized as follows. Section 2 presents the state of the art. Section 3 describes the main differences between Spanish and LSE. These differences will help to understand the main causes of the translation errors. Section 4 presents the details of the corpus used in this work. Section 5 describes briefly the two statistical translation methods used in this work, the baseline results for both systems and an error analysis. Section 6 describes the preprocessing module and section 7 presents an alternative based on factored models. Finally, the evaluation of the different alternatives is presented in section 8, and the main conclusions are summarized in section 9.

2. State of the Art

There are different sign languages depending on the country or even in every region within the country. Professor William Stokoe (Stokoe, 1960) presented the first conclusions from several studies on ASL (American Sign Language). After these studies, studies into sign languages began to increase in the USA (Anderson, 1979; Christopoulos and Bonvillian, 1985), Europe (Hansen, 1981; Frokjaer-Jensen, 1984; Notoya, 1986), Africa (Penn, Lewis and Greenstein, 1984) and Japan (Notoya, Suzuki, Furukawa and Umeda, 1986). In Spain, during the last twenty years, there have been several proposals for normalizing Spanish Sign Language (LSE: Lengua de Signos Española). M^{a} Ángeles Rodríguez (Rodríguez, 1991) carried out a detailed analysis of LSE illustrating the main characteristics. She detailed the differences between the sign language used by Deaf people and the standardization proposals. In 2007, the Spanish Government accepted the Spanish Sign Language (Lengua de Signos Española: LSE) as one of the official languages in Spain, defining a plan to invest in resources in this language, in an attempt to document and to extend this language over the entire Deaf community. In 2009, the first grammar description for LSE was presented (Herrero, 2009).

In recent years, there have been several research projects related to automatic language translation in Europe (C-Star, ATR. Vermobil, Eutrans, LC-Star, PF-Star and, finally, TC-STAR and EuroMatrixPlus) and in the USA (GALE). As regards the evaluation campaigns organized by NIST (National Institute of Standards and Technology) in the USA and the EuroMatrixPlux FP7 project in EU, the best performing translation systems are based on various types of statistical approaches (Och and Ney, 2002; Mariño et al., 2006), including example-based methods (Sumita et al., 2003), finite-state transducers (Casacuberta and Vidal, 2004) and other data driven approaches. The progress achieved over the last 10 years is down to several factors such as efficient algorithms for training (Och and Ney, 2003), context dependent models (Zens et al., 2002), efficient algorithms for generation (Koehn, 2003), more powerful computers and bigger parallel corpora, and automatic error measurements (Papineni et al., 2002; Banerjee and Lavie, 2005; Agarwal and Lavie, 2008). Another significant effort in machine translation has been the organization of several Workshops on Statistical Machine Translation (SMT). It is possible to obtain all the information on these events on the webpage http://www.statmt.org/. As a result of these workshops, there are two opensource machine translation systems called Moses (http://www.statmt.org/moses/) and Joshua (http://cs.jhu.edu/~ccb/joshua/). Moses is a phrase-based statistical machine translation system that allows machine translation system models to be built for any language pair, using a collection of translated texts (parallel corpus). On the other hand, Joshua uses synchronous context free grammars (SCFG) for statistical machine translation.

In recent years, several groups have developed prototypes for translating Spoken language into Sign Language: example-based (Morrissey, 2008), rule-based (Marshall and Sáfár, 2005; San-Segundo et al. 2008), full sentence (Cox et al., 2002) or statistical approaches (Stein et al., 2006; Morrissey et al., 2007; SiSi system; Vendrame et al., 2010) approaches. Table 1 summarizes the main characteristics of existing speech to sign language translation systems, including the current work presented in this paper.

Reference	Translation technology	Sign Language	Translation Performance	Database size (#sentences)
Cox et al., 2002	Full sentence: the system only recognizes a reduced number of already pre-translated sentences	British Sign Language (BSL)	Not reported	< 1,000
Marshall and Sáfár, 2005	HPSG based on semantic representation	BSL	Not reported	< 1,000
Bungeroth and Ney, 2004	Phrase-based model	German Sign Language (DGS)	Translation rate < 50%	< 1,500
Morrissey and Way, 2008	Example-based	Irish Sign Language (ISL)	BLEU > 50%	< 1,000
SiSi system (http://mqtt.org/projects/sisi)	Phrase-based model	BSL	Not reported	Not reported
Morrissey et al., 2007	Example-based and Phrase-based	ISL and DGS	BLEU > 50%	< 1,000
San-Segundo et al. 2008	Rule-based translation	Spanish Sign Language (LSE)	BLEU > 50%	< 500
Current work	**Phrase-based and Finite State Transducer**	**LSE**	**BLEU > 75%**	**4,080**

Table 1. Written language into sign language translation systems

As is shown, the work presented in this paper describes experiments with a relevant database, considering the small amount of data available for research into sign languages. The system presented in this paper demonstrates a very good performance compared to similar systems previously developed. The presented results are also the best results for translating Spanish into LSE using the biggest database that includes these languages. As will be presented, the preprocessing module proposed in this paper is a good approach for dealing with data sparseness when developing a speech into sign language translation system.

Given the sparseness of data for research into sign languages, in the last five years, several projects have started to generate greater resources. One of the most ambitious ones is focused on generating a corpus made up of more than 300 hours from 100 speakers in Australian Sign Language (Johnston, 2008). The RWTH-BOSTON-400 Database that contains 843 sentences with about 400 different signs from 5 speakers in American Sign Language with English annotations (Dreuw et al., 2008). The British Sign Language Corpus Project aims to create a machine-readable digital corpus of spontaneous and elicited British Sign Language (BSL) collected from deaf native signers and early learners across the United Kingdom (Schembri, 2008). And a corpus developed at The Institute for Language and Speech Processing (ILSP) which contains parts of free signing narration, as well as a considerable amount of grouped signed phrases and sentence level utterances (Efthimiou and Fotinea, 2008). There are other examples in ISL (Irish Sign Language) (Morrissey et al., 2010), NGS (German Sign Language) (Hanke et al., 2010), and Italian Sign Language (Geraci et al., 2010). In Europe, the two main research projects involving sign languages are DICTA-SIGN (http://www.dictasign.eu/) (Hanke et al., 2010; Efthimiou et al., 2010) and SIGN-SPEAK (http://www.signspeak.eu/) (Dreuw et al., 2010a and 2010b), both financed by The European Commission within the Seventh Framework. In these projects, there are important tasks for generating language resources in several European sign languages. For LSE, the biggest database was generated two years ago in a Plan Avanza project (www.traduccionvozlse.es) (San-Segundo et al., 2010). This corpus is one of the largest corpora for machine translation research involving a sign language. This corpus has been used in this work and its detailed description can be found in section 4. Not only the data but also new practice (Forster et al., 2010) and new uses of traditional annotation tools (Crasborn et al., 2010) have been developed.

3. Differences between Spanish and LSE

Spanish Sign Language (LSE), just like other sign languages, has a visual-gestural channel, but it also has grammatical characteristics similar to written languages. In linguistic terms, sign languages are as complex as written languages, despite the common misconception that they are a “simplification” of written languages. Sign languages are not mime: signs do not necessarily have a relationship to a word. They have more iconicity than spoken languages. Like spoken languages, sign language transforms meaningless units (phonemes) into units with semantic information. These phonemes of a sign are the hand shape, the palm orientation, the place of articulation, the movement and the face expressions (non-manual marks) (Lidell and Johnson, 1989; Stokoe et al., 1995).

One important difference between spoken/written languages and sign languages is sequentiality. Phonemes in spoken languages are produced in a sequence. On the other hand, sign languages have a large non-sequential component, because fingers, hands and face movements can be involved in a sign simultaneously, even two hands moving in different directions. These features give a complexity to sign languages that traditional written languages do not have. This fact makes it very difficult to write sign languages. Traditionally, signs have been represented using words (in capital letters) in Spanish (or English in the case of BSL, British Sign Language) with a similar meaning to the sign meaning. They are called glosses (i.e. ‘CAR’ for the sign ‘car’). In the last 20 years, several alternatives for writing a sign language, based on specific characteristic of the signs, have appeared in the international community: HamNoSys (Prillwitz et al., 1989), SEA (Sistema de Escritura Alfabética) (Herrero, 2004) and SignWriting (http://www.signwriting.org/). These notations allow every component of each sign to be represented (handshape, orientation, location, movement and non-manual markers).

In this work, glosses have been considered for representing signs because it is the most familiar and extended alternative in the Spanish Deaf Association. Previous works (Pizzuto et al., 2006) have point out the limitations of using glossing: problems with the discrepancies using glosses and misrepresentation of the structure of individual signs and signed discourse. In order to avoid these limitations, in this work, the sign language sentences have been generated by two LSE experts in parallel, and glossing includes additional elements such as non-speech indicators (i.e. PAY or PAY? if the sign is localized at the end of an interrogative sentence) or finger spelling indicators (i.e. DL-PETER that must be represented letter by letter P-E-T-E-R).

Another criticism (Pizzuto et al., 2006), when using glosses for developing machine translation systems, is that the system is translating from the spoken language into a constrained version of the spoken language where there is a significant overlap. Using Spanish capitalized words (glosses) for representing signs can produce the idea that just capitalizing the spoken language words, it is possible to get a first version of the LSE sentences in glosses, but it is not true in this work. Glossing, a part from the different order, includes many other signs for representing several aspects of the discourse that they do not appear in the original sentence. In order to illustrate this aspect, authors have carried out an experiment consisting of capitalizing the spoken language words and using that as the translation output. The obtained BLEU score has been 4.3%, very far from the baseline experiment (73.8% BLEU) that it will be presented in section 5.3. In order to avoid the influence of the high variability of verb conjugations in Spanish, a second experiment was carried out replacing every verb with the infinitive. The BLEU increased to 17.5%, but it is still far from the baseline experiment.

LSE has some characteristics that differ from Spanish. One important difference is the order of arguments in sentences: LSE has a **SOV** (subject-object-verb) order in contrast to **SVO** (subject-verb-object) Spanish order. An example that illustrates this behavior is shown below:

Spanish: Juan ha comprado las entradas (Juan has bought the tickets) **LSE**: JUAN ENTRADAS COMPRAR (JUAN TICKETS TO-BUY)

Comparing these two different orders in predication, the next typological differences (Table 2) can be extracted (Herrero, 2009):

SPANISH	**LSE**
Prepositions *cerca de casa (close to home)*	**Postpositions** *CASA CERCA (HOME CLOSE)*
Demonstrative + Name *ese hombre (this man)*	**Name + Demonstrative** *HOMBRE ESE (MAN THIS)*
Name + Genitive *madre de Juan (Juan's mother)*	**Genitive + Name** *JUAN MADRE (JUAN MOTHER)*
Initial interrogative particle *¿dónde está el libro? (Where is the book?)*	**Final interrogative particle** *LIBRO DÓNDE? (BOOK WHERE?)*
Auxiliary verb + Principal verb *debes comer (you must eat)*	**Principal verb + Auxiliary verb** *COMER DEBER (EAT MUST)*
Negative particle + Verb *no trabajo (I do not work)*	**Verb + Negative particle** *TRABAJAR NO (TO-WORK NO)*

Table 2. Typological differences that are related to predication order between LSE and Spanish

There are other typological differences that are not related to predication order:

- Spanish has an informative style (without topics) and LSE has a communicative style (with topics).
- Gender is not usually specified in LSE, in contrast to Spanish.
- In LSE, there can be concordances between verbs and subject, receiver or object and even subject and receiver, but in Spanish there is a concordance between verb and subject:

Subject	**Receiver**	**Subject and receiver**
Spanish: La televisión se apaga (the TV is switched off) **LSE**: TELEVISION APAGAR (TV TO-SWITCH-OFF)	**Spanish**: Yo le respeto (I respect him) **LSE**: YO RESPETAR-a-él (I RESPECT-HIM)	**Spanish**: Te explica (he explains to you) **LSE**: EXPLICAR-él-a-ti (EXPLAIN-HIM-TO-YOU)

- The use of classifiers is common in LSE, but they are not in Spanish. A classifier is a manual configuration that substitutes a class of objects, considering the Herrero's definition (Herrero, 2009). For example:

Spanish: debe acercarse a la cámara (you must approach the camera)
LSE: FOTO CLD_GRANDE_NO CLI_ACERCARSE DEBER (PHOTO CLD_BIG_NO CLI_APPROACH MUST)[1]

- Articles are used in Spanish, but not in LSE. For example:

Spanish: La televisión se apaga (The TV is switched off)
LSE: TELEVISION APAGAR (TV TO-SWITCH-OFF)

- Plural is descriptive in LSE, but not in Spanish. The way the flowers sign is represented provides information about how the flowers are situated. For example:

Spanish: flor (flower) **LSE**: FLOR (FLOWER)	**Spanish**: flores (flowers) **LSE**: CL-"flores" (CL-"flowers")[2]

- There is a difference between an absent and present third person in LSE, but there is no absent third person in Spanish.
- In LSE, there is the possibility of using double reference, not in Spanish.
- LSE is a language with ample flexibility, and homonymy between substantive and adjective is usual, so most nouns can be adjectives and vice versa. But there are few cases in Spanish.
- In Spanish, there is a copula in non-verbal predications (the verb 'to be', *ser* and *estar* in Spanish), but there is not in LSE (except some locative predications). For example:

Spanish: Antonio está en la Universidad (Antonio is in the University)
LSE: ANTONIO/UNIVERSIDAD ALLÍ (ANTONIO/UNIVERSITY THERE)

- There is a difference between inclusive and exclusive quantifiers in LSE, but not in Spanish.
- There are Spanish impersonal sentences with "*se*" pronoun, but not in LSE. For example:

Spanish: Se come bien (you eat well)
LSE: COMER BIEN (TO-EAT WELL)

- It is important to comment that LSE is more lexically flexible than Spanish, and it is perfect for generating periphrasis through its descriptive nature and because of this, LSE has fewer nouns than Spanish.
- LSE has fewer signs per sentence (5.2 in our database) than Spanish (7 in our database).
- LSE has smaller vocabulary variability. LSE has a vocabulary of around 10,000 signs (Pinedo, 2000) while Spanish has several million different words. Good examples are the different verb conjugations.

4. Parallel corpus

In order to develop a translation system focused on the domain of the renewal of an Identity Document (ID) and Driver's license (DL), a database, including a parallel corpus, is necessary. This database has been obtained with the collaboration of Local Government Offices where the aforementioned services (ID and DL) are provided (San-Segundo et al., 2010). For a period of three weeks, the most frequent explanations (from government employees) and the most frequent questions (from the user) were taken down and more than 5,000 sentences were noted. These 5,000 sentences were analyzed because not all of them refer to ID or DL, so sentences were selected manually in order to develop a system in a specific domain. Finally, 4,080 sentences were collected.

These sentences were translated into LSE, both in text (sequence of signs) and in video, and compiled in an excel file. The translation was carried out by two LSE experts in parallel. When there was any discrepancy between them, a committee of four people who knew LSE took the decision: select one of the LSE expert proposals, propose a new one translation alternative, or considering both proposals as alternative translations. The committee was made up of one Spanish linguist, two Deaf LSE experts

[1] CLD: means a Descriptive CLassifier and CLI means an Instrumental CLassifier.

[2] CL: means a generic CLassifier

and Spanish linguist expert on LSE. The excel file contains eight different information fields: “ÍNDICE” (sentence index), “DOMINIO” (domain: ID or DL renewal), “VENTANILLA” (window where the sentence was collected), “SERVICIO” (service provided when the sentence was collected), if the sentence was pronounced by the government employee or user, sentence in Spanish (CASTELLANO), sentence in LSE (sequence of signs), and link to the video file with LSE representation. The main features of the corpus are summarized in Table 3. These features are divided depending on the domain (ID or DL renewal) and whether the sentence was spoken by the government employee or the user.

	ID		**DL**	
Government employee	**Spanish**	**LSE**	**Spanish**	**LSE**
Sentence pairs	1,425		1,641	
Different sentences	1,236	389	1,413	199
Words or signs per sentence	5.9	4.4	10.4	7.8
Running words	8,490	6,282	17,113	12,741
Vocabulary	652	364	527	237
User	**Spanish**	**LSE**	**Spanish**	**LSE**
Sentence pairs	531		483	
Different sentences	458	139	389	93
Words or signs per sentence	5.2	3.7	6.5	4.7
Running words	2,768	1,950	3,130	2,283
Vocabulary	422	165	294	133

Table 3. Main statistics of the corpus

For the system development, two types of files were generated from the database: text files and sign files. Text files are made up of Spanish sentences of the parallel corpus and sign files contain their LSE translations (LSE sentences are sign sequences).

The corpus was divided randomly into three sets: training (75%), development (12.5%) and test (12.5%), carrying out a 8-step cross-validation process. So, eight sets of parallel corpus (of Spanish and LSE sentences) were created: six of them were used for training, one for development and the last one for testing. For each experiment presented in this paper, the results show the average of this cross-validation process.

5. Statistical translation strategies

In this paper, two different statistical strategies have been considered: a phrase-based system and a Statistical Finite State Transducer. The preprocessing module has been evaluated with both translation strategies. This section describes briefly the architectures used for the experiments.

5.1 Phrase-based translation system

The Phrase-based translation system is based on the software released at the 2009 NAACL Workshop on Statistical Machine Translation (http://www.statmt.org/wmt09/)(Figure 2).

The phrase model has been trained by following three steps. The first one is word alignment computation using GIZA++ (Och and Ney, 2003). GIZA++ is a statistical machine translation toolkit that is used to train IBM Models 1-5 and an HMM word alignment model. This package also contains the source for the ‘mkcls’ tool which generates the word classes necessary for training some of the alignment models. In this case, GIZA++ has been used to calculate the alignments between words and signs in both directions (Spanish-LSE and LSE-Spanish). To generate the translation model, the parameter “alignment” was fixed to “target-source” as the best option, based on experiments on the development set. In this work, only this target-source alignment was considered (LSE-Spanish). In this configuration, alignment is guided by signs: this means that in every sentence pair alignment, each word can be aligned to one or several signs (but not the opposite). It is also possible that some words were not aligned to any sign. When combining the alignment points from all sentence pairs in the training set, it is possible to have all possible alignments: several words aligned to several signs.

The second step is phrase extraction (Koehn et al. 2003). All phrase pairs that are consistent with the word alignment (target-source alignment in our case) are collected. The maximum size of a phrase has been fixed at 7 based on development experiments over the development set (see previous section).

Finally, the last step is phrase scoring. In this step, the translation probabilities are computed for all phrase pairs. Both translation probabilities are calculated: forward and backward.

The Moses decoder is used for the translation process (Koehn, 2010). This program is a beam search decoder for phrase-based statistical machine translation models. In order to obtain a 3-gram language model, the SRI language modeling toolkit has been used (Stolcke, 2002).

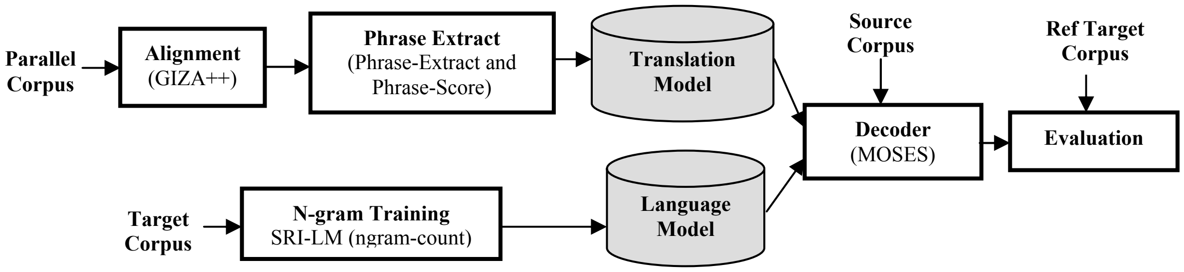

Figure 2. Phrase-based translation architecture.

5.2. Statistical Finite State Transducer

The translation based on SFST is carried out as set out in Figure 3.

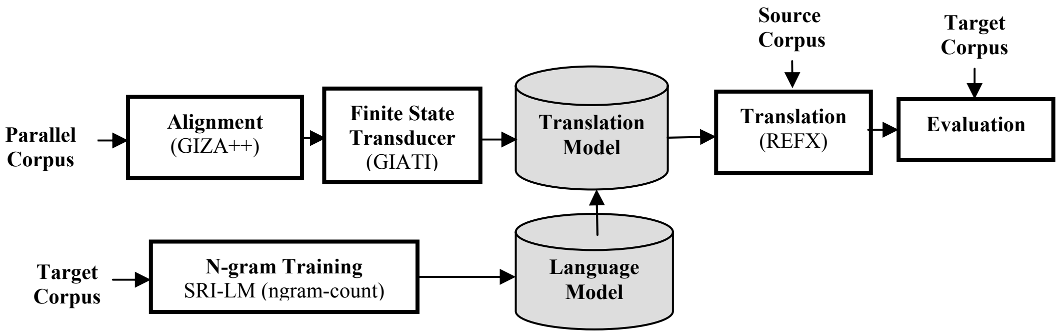

Figure 3. Diagram of the SFST-based translation module

The translation model consists of an SFST made up of aggregations: subsequences of aligned source and target words. The SFST is inferred from the word alignment (obtained with GIZA++) using the GIATI (Grammatical Inference and Alignments for Transducer Inference) algorithm (Casacuberta and Vidal, 2004). The SFST probabilities are also trained from aligned corpora. The software used in this paper has been downloaded from http://prhlt.iti.es/content.php?page=software.php.

5.3. Baseline results and error analysis

For the experiments, the corpus (described in section 4) was divided randomly into three sets: training (75%), development (12.5%) and test (12.5%), carrying out a cross-validation process. The results presented are the average of the 8 steps (as explained in section 4). The training set was used to train the translation and language models, and the development sets were used for adjusting the weights, tuning some parameters and the analysis of the probability threshold (in the case of computing the word-tag list automatically) as presented in the next section.

The baseline for both systems (phrase-based and SFST) consists of not considering any preprocessing module before the translation module. So the translation models must be trained with raw source and a target corpus: without any type of preprocessing or factorization, i.e., Spanish sentences contain words and LSE sentences contain signs from the original database. For example:

> Spanish: *en esta hoja viene todo lo necesario (In this paper, you have all you need)*
> LSE: *PAPEL ESTE INFORMACIÓN DETALLE TODO (PAPER, THIS INFORMATION DETAIL ALL)*

For evaluating the performance of the translation systems, different accuracy metrics are presented: BLEU (BiLingual Evaluation Understudy)(Papineni et al., 2002) in percentage and NIST. The NIST metric is based on the BLEU metric, but with some alterations (Doddington, 2002). It was proposed by the National Institute of Standards and Technology in the USA. BLEU and NIST measurements have been computed using the NIST tool (mteval.pl). Two error metrics have also been also added to the results: mSER (multiple reference Sign Error Rate) and PER (multiple reference Position independent sign Error Rate). It is important to note that BLEU and NIST are accuracy metrics while mSER and PER are error metrics. In order to analyze the significance of the differences between several systems, for every BLEU result, the confidence interval (at 95%) is also presented. This interval is calculated by using the following formula:

$$\pm\Delta = 1{,}96\sqrt{\frac{BLEU(100-BLEU)}{n}} \qquad (1)$$

n is the number of signs used in evaluation, in this case n=23,256.

Baseline systems	BLEU (%)	±Δ	NIST	mSER(%)	PER(%)
Phrase-based approach	**73.8**	**0.57**	**8.98**	**26.77**	**19.38**
SFST-based approach	**70.6**	**0.59**	**8.49**	**26.36**	**23.84**

Table 4. Results for the baseline systems: phrase-based and SFST-based systems

Table 4 presents the results for the baseline systems. As is shown, the phrase-based system has better performance that the SFST. As regards the confidence interval (±Δ), this difference is statistically significant because there is no overlap between the confidence intervals from different systems. These results are very good compared to an open vocabulary translation system because the system described in this paper is focused on a restricted domain and all the sentences in the database are included in this domain: there is no sentence out of the domain.

Throughout the paper, BLEU will be considered as the main evaluation metric for comparing different translation systems because BLEU presents a very good correlation to human evaluations (Papineni et al., 2002). Section 8.3 presents the results of a human evaluation showing that BLEU is the automatic metric that has the best correlation to human judgments also in this work.

In order to analyze the best strategy for improving these results, an error analysis has been carried out to try to establish a relationship between these errors and the main differences between Spanish and LSE (section 3). Table 5 presents this analysis.

Rate	Error description	Main Causes
47.8%	The most important type of error is related to the fact that in Spanish there are more words than signs in LSE. This circumstance provokes different types of errors: - Generation of many phrases in the same output, producing a high number of insertions. - When dealing with long sentences there is the risk that the translation model cannot deal properly with the big distortion. This produces important changes in order and sometimes the sentence is truncated producing several deletions.	The most important causes are: - Long periphrasis in Spanish (26.0%) - Articles and determiners (21.0%) - Others like different use of gender, plural or copula verbs (0.8%)
17.4%	Secondly, there are several ordering errors related to the different orders in predication: LSE has a SOV (Subject-Object-Verb) while Spanish SVO (Subject-Verb-Object).	Different order in predication (17.4%)
15.0%	Wrong generation of the different classifiers needed in this sentence.	The common use of classifiers in LSE and their absence in Spanish
14.5%	Out Of Vocabulary words (OOV). In the evaluation set there are words that were not in the training set, so the translation and language models cannot deal properly with these words.	When translating Spanish into LSE, there is a relevant number of OOVs due to the higher variability presented in Spanish. For example, the verb conjugations. In Spanish there are many verb conjugations that are translated into the same sign sequence. So, when one of these conjugations appears in the evaluation set provokes an OOV error.
5.3%	Finally, there are some deletions when translating very specific names, even when they are in the training set.	In this case, these errors come from the need to generate a periphrasis in LSE that the system cannot generate properly, provoking some deletions

Table 5. Error Analysis for both baseline systems: phrase-based and SFST-based systems

As is shown, the main causes of the translation errors are related to the different variability in the vocabulary for Spanish and LSE (much higher in Spanish), the different number for words or signs in the sentences (higher in Spanish) and the different predication order. In order to improve these results, this paper proposes to introduce a preprocessing module able to reduce the Spanish variability and remove non-relevant Spanish words from the source sentence. This preprocessing will consist of a categorization of the sequence of words in the source sentence (Spanish).

6. Preprocessing module

As was presented in Figure 1, the preprocessing module proposed in this paper analyzes the source language sentence (sentence in Spanish) and replaces Spanish words with their associated tags. A tag refers to a word sequence connected by dash characters in capital letter. This word sequence provides syntactic-semantic information about every Spanish word. Figure 4 shows an extract of the word-tag list. This list is composed of Spanish words and their corresponding tags, including the English translation in parenthesis.

word TAG (word and tag in English)
…
cerrado CERRAR-YA (*closed CLOSE-ALREADY*)
cerramos CERRAR (*we close CLOSE*)
cerrar CERRAR (*to close CLOSE*)
cobradas COBRAR-YA (*charged CHARGE-ALREADY*)
cobro COBRAR (*I charge CHARGE*)
coge COGER (*you get GET*)
cogido COGER-YA (*gotten GET-ALREADY*)
coja COGER (*you get GET*)
…

Figure 4. Extract of the word-tag list.

This module uses a list of Spanish words (the vocabulary in this restricted domain) and the corresponding tags, previously computed during the training process. For every word, only one tag is associated. So given a Spanish sentence, this module implements a simple algorithm: for all words in the sentence, the preprocessing module looks for this word in the list and replaces it with the associated tag. It is important to comment on two main aspects. The first one is that there is a tag named "non-relevant" associated to those words that are not useful for translating the sentence. The second one is that if the Spanish word is not on the list (it is an Out Of Vocabulary word), this word is not replaced with any tag: this word is kept as it is.

Finally, it is important to remark that in order to train the statistical translation module when using the preprocessing module, it is necessary to retrain the translation models considering the tagged source language, not the original word sentences. This way, the translation models learn the relationships between tags and signs.

6.1. Word-tag list generation: Manual vs. Automatic

The main issue for implementing the preprocessing module is to generate the list of the Spanish words with the associated tags. In this section, two methods will be presented: one manual and another automatic. For the manual approach, the preprocessing module considers the categories used in the rule-based translation system previously developed for this application domain (San-Segundo et al., 2008). Every category is an identifier that provides syntactic and/or semantic information about the word. In this case, the natural language translation module was implemented using a rule-based technique considering a bottom-up strategy. The translation process is carried out in two steps. In the first one, every word is mapped onto one syntactic-pragmatic tag. After that, the translation module applies different rules that convert the tagged words into signs by means of grouping concepts or signs and defining new signs. These rules can define short and large scope relationships between the concepts or signs. For the manual approach, the list of words and categories used in the first step (1,014 categories manually generated) has been considered for these experiments. Generating this list manually is a subjective, slow and difficult task, which is why the main proposal in this paper is to define a procedure for calculating this list automatically.

In order to obtain the tags automatically, the main idea consists of tagging every Spanish word with the sign that shows the highest probability of being the translation of this word. This probability is calculated by using the lexical model obtained from the word-sign GIZA++ alignments. This lexical model indicates several potential signs that might be the translation of each word with some translation probability. For example:

CANTIDAD abonar 0.0526316
DINERO abonar 0.0526316
EUROS abonar 0.0454545
NULL abonar 0.1578947
PAGAR abonar 0.2631579
PRECIO abonar 0.4210526
YO abonar 0.0526316

Spanish word "*abonar*" can be translated by signs "*CANTIDAD*", "*DINERO*", "*EUROS*", "*NULL*", "*PAGAR*", "*PRECIO*" and "*YO*" with their probabilities. *NULL* indicates that this word is not aligned to any sign in this corpus with a certain probability.

Given the lexical model, the proposed automatic approach consists of assigning the sign with the highest probability of being the translation of this word to every Spanish word. But this tag is assigned only if this probability is higher than a threshold. If there is no probability higher than the threshold, the tag for this word will be the same word: the word is kept as it is. If the most probable sign is "*NULL*" and its probability is higher than this threshold, this word will tagged with the "non-relevant" tag.

In order to generate the words-tags file automatically, it is necessary to define a probability threshold. Using the development sets, several experiments were carried out with different thresholds between 0.0 and 1.0 probability. Figure 5 shows the results (BLEU) with each probability threshold considering the both phrase-based and SFST-based translation strategies.

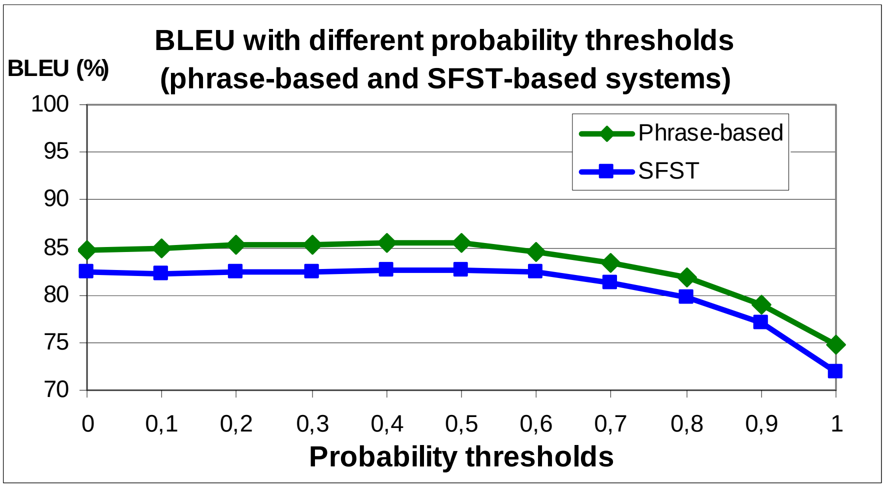

Figure 5. Results of the phrase-based system and SFST when using different probability thresholds for generating automatically the word-tag mapping

As is shown, there is a stable behavior between 0.2 and 0.5, so for the experiments the threshold was fixed to 0.4 for both translation strategies. For these experiments, the confidence interval (at 95%) ±Δ is less than 0.60%. The differences in BLEU between a 1.0 threshold and a 0.4 threshold are higher than 10% absolute, so they are statistically significant: there is no overlapping between the confidence intervals.

6.2. Dealing with non-relevant words

When implementing the preprocessing module, several strategies for dealing with the "non-relevant" words have been proposed:

- In the first alternative, all the words are replaced by their tags with the exception of those words that they do not appear in the list (OOV words). As was commented on before, they are kept as they are. In the word-tag list, there is a "non-relevant" tag mapped to words that are not relevant for the translation process (named "basura" (garbage)). This alternative will be referred to in the experiments like "**Using tags**". For example:

  Original source sentence: *en esta hoja viene todo lo necesario (In this paper, you have all you need)*

  Categorized source sentence: *basura ESTE PAPEL basura basura basura NECESARIO*

  Target sentence: *PAPEL ESTE INFORMACIÓN DETALLE TODO (PAPER, THIS INFORMATION DETAIL ALL)*

- The second proposed alternative was not to tag any word in the source language but to remove non-relevant words from the source lexicon (associated to the "non-relevant" tag). This alternative will be referred to in the experiments as "**Removing non-relevant words from the source lexicon**". For example:

  Original source sentence: *en esta hoja viene todo lo necesario*

  Categorized source sentence: *esta hoja necesario*

  Target sentence: *PAPEL ESTE INFORMACIÓN DETALLE TODO*

- Finally, the third alternative proposes to replace words with tags (with the exception of OOVs) and to remove "non-relevant" tags. This alternative will be referred to in the experiments as "**Using tags and removing non-relevant tags**". For example:

  Original source sentence: *en esta hoja viene todo lo necesario*

  Categorized source sentence: *ESTE PAPEL NECESARIO*

  Target sentence: *PAPEL ESTE INFORMACIÓN DETALLE TODO*

In the experiments (section 8), the three alternatives will be evaluated and discussed.

7. Factored models

For the phrase-based translation strategy, there is the possibility to train factored models in order to include this information in the translation process (Koehn, 2007). This possibility is an extension of phrase-based statistical machine translation models that enables the straightforward integration of additional annotations at the word-level (linguistic markup or automatically generated word classes). The main idea is to add additional annotation at the word level. A word in this framework is not only a token, but a vector of factors that represents different levels of annotation. The translation of factored representations of input words into the factored representations of output words is broken up into a sequence of mapping steps that either translates input factors into output factors, or generates additional output factors from existing output factors. The information included in these factored models can be a tag with semantic information, sort of word (name, article, verb, adverb, preposition, etc.), gender or number of word, verb tense, adverb characteristics, etc.

Each word in the corpus becomes a vector with the following format: Word|Factor1|Factor2|... For example, word "documentación" becomes: "documentación|DOCUMENTACIÓN|nombre|singular".

As an alternative to the preprocessing module, this paper also analyses the possibility of integrating the tagging information using factored models, but only for the phrase-based translation system. In this work, only the source language has been factored with an additional factor, its tag.

Like the three different preprocessing strategies presented in section 6.2, three different alternatives has been considered when generating the factored models.

- In the first alternative, all the words in the source language are factored and several translation models are trained (word-sign and tag-sign). Only two factors have been considered: word and tag. For example:

  Original source sentence: *en esta hoja viene todo lo necesario (In this paper, you have all you need)*

  Factorized source sentence: en|*basura esta*|*ESTE hoja*|*PAPEL viene*|*basura todo*|*basura lo*|*basura necesario*|*NECESARIO*

  Target sentence: *PAPEL ESTE INFORMACIÓN DETALLE TODO (PAPER, THIS INFORMATION DETAIL ALL)*

- The second proposed alternative was to keep the original words (without additional factors), but removing non-relevant words from the source lexicon.

  Original source sentence: *en esta hoja viene todo lo necesario*

  Factorized source sentence: *esta hoja necesario*

  Target sentence: *PAPEL ESTE INFORMACIÓN DETALLE TODO*

- Finally, in third alternative all the words are factored and "non-relevant" words are removed.

  Original source sentence: *en esta hoja viene todo lo necesario*

Factorized source sentence: *esta|ESTE hoja|PAPEL necesario|NECESARIO*

Target sentence: *PAPEL ESTE INFORMACIÓN DETALLE TODO*

8. Experiments and discussion

Like the baseline experiments, for these experiments the corpus was divided randomly in training (75%), development (12.5%) and test (12.5%), and by carrying out a cross-validation process made up of 8 steps. The results presented are the average of the 8 steps. The results are compared with the baseline systems presented in section 5.2. The evaluation metrics are also BLEU in percentage, NIST, mSER and PER. Each BLEU result shows also a 95% confidence interval in order to see whether the result is statistically significant. This interval is calculated by using equation 1.

8.1. Factored models vs. preprocessing

This section presents the results for comparing the factored models and the preprocessing proposed in this work. This comparison is made by using the phrase-based translation together with the manual method for generating the word-tag list. Table 6 includes all the evaluation metrics for the three alternatives considered in previous sections for dealing with non-relevant words.

Translation system		BLEU(%)	±Δ	NIST	mSER(%)	PER(%)
Baseline		**73.8**	**0.56**	**8.98**	**26.77**	**19.38**
Using tags	Factored models	72.7	0.57	8.85	26.84	21.54
	Preprocessing	80.6	0.51	9.55	18.23	14.44
Removing non-relevant words from the source lexicon	Factored models	77.1	0.54	9.21	22.08	17.54
	Preprocessing	77.1	0.54	9.21	22.08	17.54
Using tags and removing “non-relevant” tags	Factored models	78.8	0.53	9.35	20.84	17.80
	Preprocessing	**83.5**	**0.48**	**9.80**	**15.37**	**12.54**

Table 6. Comparison between factored models and including the preprocessing module computing the word-tag list manually.

The first analysis is that factored models proposed in the phrase-based architecture perform worse than the preprocessing methods proposed. In factored models, two translation models were generated: word-sign and tag-sign, but it seems that the process for tuning the weights of these models does not reach the optimum ones. However, in both cases (factored model and preprocessing); removing words tagged with the “non-relevant” tag improves the results significantly: the differences between the different systems are greater than the confidence intervals (±Δ). This aspect depends a lot on the relationship between the two languages considered in this study. LSE uses fewer signs than words in Spanish, so when removing non-relevant words (such as articles), the translation model has less dispersion and the results improve. Using the preprocessing module, the system increases its performance from BLEU 73.8% to 83.5% (13.1% relative increment). This improvement is statistically significant compared to the 95% confidence interval computed for these cases. In the next section, a detailed analysis of the translation errors will be presented.

8.2. Comparing manual and automatic word-tag list generation methods

Table 7 presents the results of comparing manual and automatic methods for obtaining the word-tag list used by the preprocessing module.

Phrase-based translation System		**BLEU(%)**	**±Δ**	**NIST**	**mSER(%)**	**PER(%)**
Baseline		**73.8**	**0.56**	**8.98**	**26.77**	**19.38**
Using tags	Automatic	79.2	0.52	9.33	19.61	16.28
	Manual	80.6	0.51	9.55	18.23	14.44
Removing non-relevant words from the source lexicon	Automatic	76.7	0.54	9.24	21.58	18.17
	Manual	77.1	0.54	9.21	22.08	17.54
Using tags and removing “non-relevant” tags	Automatic	**81.0**	**0.50**	**9.56**	**17.70**	**15.05**
	Manual	**83.5**	**0.48**	**9.80**	**15.37**	**12.54**

Table 7. Comparing automatic vs. manual word-tag list generation used by the preprocessing module incorporated in the phrase-based translation system.

With the preprocessing module, the BLEU has increased from 73.8% to 81.0% or 83.5% as regards automatic or manual word-tag list generation methods respectively. These figures show 9.8% and 13.3% relative improvements. As is shown, the results obtained with the automatic word-tag-list generation method are statistically worse than those obtained with the manual one (the difference is higher than the confidence interval), but they are very close and the effort required for generating automatic tags is considerably lower.

By analyzing the results for the different strategies when dealing with non-relevant words, it is shown that replacing words with tags and removing “non-relevant” words are complementary actions that, when considered together, allow the best results for phrase-based and SFST-based translation systems to be achieved.

Table 8 shows the translation results for the Statistical Finite State Transducer, comparing manual and automatic word-tag list generation. In this architecture, the factored models were not considered.

SFST-based translation System		**BLEU(%)**	**±Δ**	**NIST**	**mSER(%)**	**PER(%)**
Baseline		**70.6**	**0.58**	**8.49**	**26.36**	**23.84**
Using tags	Automatic	70.0	0.59	8.42	27.13	25.01
	Manual	72.8	0.57	8.67	24.37	21.73
Removing non-relevant words from the source lexicon	Automatic	76.5	0.55	9.01	21.74	19.78
	Manual	75.3	0.55	8.91	22.60	20.37
Using tags and removing “non-relevant” tags	Automatic	**78.4**	**0.53**	**9.23**	**19.97**	**17.96**
	Manual	**79.6**	**0.52**	**9.37**	**18.78**	**16.39**

Table 8. Comparing automatic vs. manual word-tag list generation used by the preprocessing module incorporated in the SFST-based translation system.

Like the phrase-based strategy, using the preprocessing module with SFST increases BLEU from 70.6% to 78.4% and 79.6% as regards automatic or manual word-tag list generation methods respectively. These figures show 11.0% and 12.7% relative improvements. These improvements are very similar to those obtained for the phrase-based system. As expected, by comparing automatic and manual methods, the results obtained with the automatic strategy are slightly worse but the confidence intervals overlap so these results are not significantly different. The best results are obtained when the words are replaced by categories and the “non-relevant” categories are also removed from the source language.

As commented on in section 6.1, the automatic method for generating the word-tag list is based on the lexical model obtained from the word-sign GIZA++ alignments. As regards this aspect, it is possible to consider an iterative process where the tags obtained in the "i" iteration have been computed based on the lexical model obtained from GIZA++ alignments between tags and target signs in the iteration "i-1". Figure 6 shows the BLEU evolution depending on the number of iterations. As is shown, with two iterations it is possible to get slightly better results, but the BLEU remains constant for further iterations.

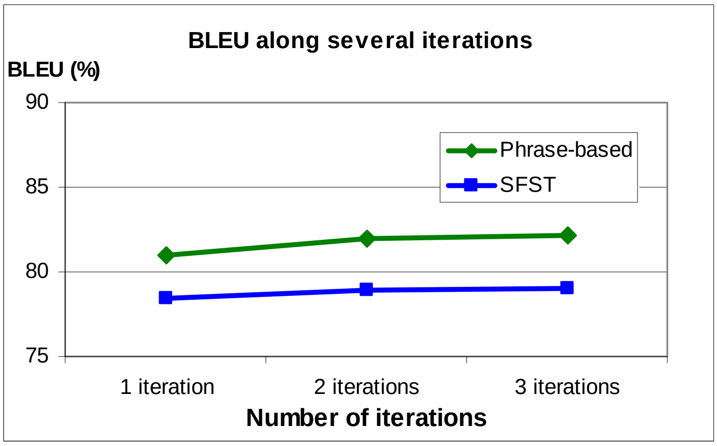

Figure 6. Evolution of BLEU for using automatic tag generation in several iterations

In order to analyze the main reasons for the improvement found with the preprocessing method proposed in this work, a new error analysis has been performed similar to that presented in section 5.3 for the baseline systems. This new analysis has considered the errors made by the systems that include the preprocessing module with these characteristics: the word-tag list has been computed automatically, and the preprocessing consists of replacing words with tags (with the exception of OOVs) and removing "non-relevant" tags. Table 5 presents this analysis.

Rate	Error description	Main Causes
33.7%	Errors because in Spanish there are more words than signs in LSE: - Generation of many phrases in the same output, producing a high number of insertions. - Big distortions produce significant changes in order and sometimes the sentence is truncated producing several deletions.	The most important causes are: - Long periphrasis in Spanish (17.5%) - Articles and determiners (15.2%) - Others like different use of gender, plural or copula verbs (1%)
22.5%	Out Of Vocabulary words.	Higher variability presented in Spanish.
19.0%	Wrong generation of the different classifiers needed in this sentence.	The common use of classifiers in LSE and their absence in Spanish
16.3%	Errors related to the different orders in predication: LSE has a SOV while Spanish SVO.	Different order in predication (16.3%)
8.5%	Finally, there are some deletions when translating very specific names, even when they are in the training set.	Periphrasis in LSE not generated properly.

Table 9. Error Analysis incorporating the preprocessing module

By comparing this analysis with that presented in Table 5, it is possible to see how errors brought about by the different variability and sentence lengths between Spanish and LSE have been reduced relatively by incorporating the preprocessing module. The errors provoked by the different predication order have also been reduced. The rest of the errors increase their percentage given that the sum of all error percentages is 100%.

In conclusion, the preprocessing module allows the variability in the source language to be reduced together with the number of tokens that make up the input sentence. These two aspects give rise to a significant reduction in the number of source-target alignments the system has to train. When having a small corpus, as is the case in many sign languages, this reduction of alignment points permits to get better training models with fewer data.

8.3 Human Evaluation

In order to complete the analysis, a human evaluation was carried out for one week. The two experts involved in the corpus generation evaluated one of the test lists translated by the two baseline systems (phrase-based and SFST) and these two systems including the preprocessing module implementing the best alternative: replacing words with tags (with the exception of OOVs) and removing "non-relevant" tags. The word-tag list has been generated automatically. The output sentences (sign sequences) were presented to the experts randomly, mixing sentences from all the evaluated systems. Both experts evaluated every sentence with one of the three possible scores:

- 1 when the sentence is well constructed and the meaning is the same as the original one.
- 0.5 when there are errors but the sentence is understandable and the meaning is the same as the original one.
- 0 when the sentence is not understandable or the meaning is not the same as the original one.

Both experts evaluated all of the sentences and the results presented in *Figure 7* are the average of both expert evaluations. From the human evaluation, it is possible to conclude that the preprocessing module improves the score significantly: there is no overlap between confidence intervals. On the other hand, the difference between the phrase-based approach and the SFST is not significant in both cases: with and without the preprocessing module.

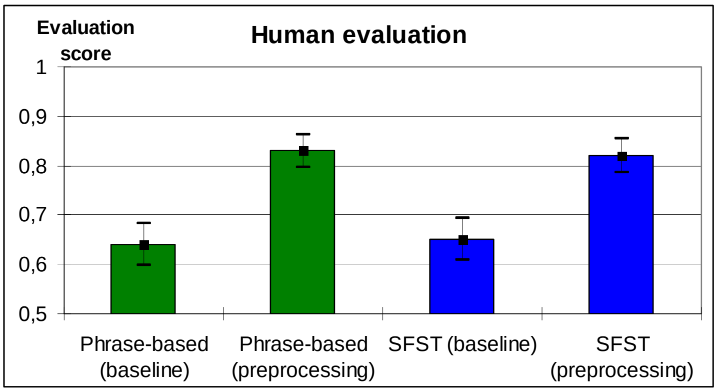

Figure 7. Human evaluation

It is important to note that this human evaluation has been carried out by two LSE experts in the sign language sentences in texts. This text representation must be passed to an avatar for presenting the signs. For carrying out a human evaluation of the whole system, Deaf people (not necessary expert in LSE) should also evaluate how the avatar represents these signs. This is an important aspect that the authors will address in the near future.

Finally, in order to complete the human evaluation, Table 10 shows the Pearson correlation between the different automatic metrics and the human evaluation. This correlation produces a number between –1 (opposite behaviours) and 1 (similar behaviours). A 0 correlation means that there is no relationship between these two metrics. This table also includes p-values for reporting the correlation significance. As is shown, BLEU is the automatic metric with the highest correlation (close to 0.9), which is why this, BLEU has been used for discussing the main results presented in this work.

Translation system	BLEU	NIST	100%-mSER	100%-PER
Phrase-based (baseline)	**0.87** (p=0.003)	0.85 (p=0.004)	0.68 (p=0.010)	0.65 (p=0.010)
Phrase-based (preprocessing)	**0.90** (p=0.002)	0.90 (p=0.002)	0.65 (p=0.013)	0.62 (p=0.014)
SFST (baseline)	**0.83** (p=0.005)	0.82 (p=0.005)	0.67 (p=0.010)	0.68 (p=0.007)
SFST (preprocessing)	**0.89** (p=0.002)	0.88 (p=0.002)	0.66 (p=0.012)	0.64 (p=0.012)

Table 10. Analysis of correlations between automatic metrics and human evaluation

9. Conclusions

This paper has described a preprocessing module for improving the performance of Spoken Spanish into a Spanish Sign Language (LSE: Lengua de Signos Española) translation system. The proposed module has been incorporated into two well-known statistical translation architectures: a phrase-based system (Moses) and a Statistical Finite State Transducer. Both architectures have been adapted to translating sentences in a specific application domain: the renewal of Identity Documents and the Driver's License.

The preprocessing module uses a word-tag list for tagging the source sentence. In this module, all the words in the input sentence are replaced by their tags with the exception of those words that do not appear in the list (OOV words). They are kept as they are. After that, the "non-relevant" tags are removed from the input sentence. The word-tag list is generated automatically using the lexical model obtained from the word-sign GIZA++ alignments. Given the lexical model, the tag associated to a given word is the sign with the highest probability of being the translation of this word. But this tag is assigned only if this probability is higher than a threshold. If there is no probability higher than the threshold, the tag for this word will be the same word. If the most probable sign is "*NULL*" and its probability is higher than this threshold, this word will tagged with the "non-relevant" tag. This probability threshold is fixed to 0.4 based on development evaluations.

By comparing the different error analysis carried out on the baseline and final systems, it is possible to conclude that the preprocessing module has permitted the variability to be reduced together with the number of tokens in the source language thus reducing the number of source-target alignments the system has to train. When having a small corpus as it is the case of many sign languages, this reduction in alignment points permits better training models with fewer data to be obtained.

The evaluation results have revealed a significant improvement in both statistical translation strategies. In the phrase-based system, the proposed categorization allowed an increase in BLEU from 73.8% to 81.0%, and for the SFST, the BLEU was increased from 70.6% to 78.4%.

This paper has also presented a human evaluation. Considering this evaluation, the preprocessing module obtains an increase from 0.64 to 0.83 in the phrase-based system and an increase from 0.65 to 0.82 in the case of the SFST.

Acknowledgements

The authors would like to thank the eSIGN (Essential Sign Language Information on Government Networks) consortium for permitting the use of the eSIGN Editor and the 3D avatar in this research work. The authors would also like to thank discussions and suggestions from the colleagues at GTH-UPM and Fundación CNSE. This work has been supported by Plan Avanza Exp Nº: TSI-020100-2010-489), INAPRA (MEC ref: DPI2010-21247-C02-02), and SD-TEAM (MEC ref: TIN2008-06856-C05-03) projects and FEDER program. The authors would also like to thank Mark Hallett for the English revision.

References

Agarwal, A., and Lavie, A., 2008. "Meteor, m-bleu and m-ter: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output", In Proceedings of the workshop on Statistical

Machine Translation at the 46th Annual Meeting of the Association of Computational Linguistics (ACL-2008), Columbus, Ohio, USA. pp 115-118.

Anderson LB., 1979 “Aspect in Sign Language Morphology: The Role of Universal Semantics and Pragmatics in Determining Grammatical categories”, Linguistics Research Laboratory, Gallaudet College (for the Symposium on Tense/Aspect: between semantics and pragmatics, UCLA, 4–6 May).

Bungeroth J., Ney, H., “Statistical Sign Language Translation”. In Proceedings of the workshop on Representation and Processing of Sign Languages, LREC 2004, Lisbon, Portugal. pp 105-108.

Casacuberta F., Vidal. E., 2004. “Machine Translation with Inferred Stochastic Finite-State Transducers”. Computational Linguistics, Vol. 30, No. 2, pp. 205-225.

Christopoulos, C. Bonvillian, J. 1985. “Sign Language, Journal of Communication Disorders” 18 (1985):1–20.

Crasborn O., Sloetjes H.. 2010. “Using ELAN for annotating sign language corpora in a team setting”. In Proceedings of the 4th workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta. pp 61-65.

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

Doddington, G. 2002 "Automatic evaluation of machine translation quality using n-gram cooccurrence statistics". In Proceedings of the Human Language Technology Conference (HLT), San Diego, CA USA, pp 128–132.

Dreuw P., Neidle C., Athitsos V., Sclaroff S., and Ney H. 2008. “Benchmark Databases for Video-Based Automatic Sign Language Recognition”. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco. pp 1115-1120.

Dreuw P., Ney H., Martinez G., Crasborn O., Piater J., Miguel Moya J., and Wheatley M., 2010a “The SignSpeak Project - Bridging the Gap Between Signers and Speakers”. In Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta. pp 73-76.

Dreuw P., Forster J., Gweth Y., Stein D., Ney H., Martinez G., Verges Llahi J., Crasborn O., Ormel E., Du W., Hoyoux T., Piater J., Moya Lazaro JM, and Wheatley M. 2010b “SignSpeak - Understanding, Recognition, and Translation of Sign Languages”. In Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta. pp 73-76.

Efthimiou E., and Fotinea, E., 2008 “GSLC: Creation and Annotation of a Greek Sign Language Corpus for HCI” LREC. Marrakech, Morocco.

Efthimiou E., Fotinea S., Hanke T., Glauert J., Bowden R., Braffort A., Collet C., Maragos P., Goudenove F. 2010. “DICTA-SIGN: Sign Language Recognition, Generation and Modelling with application in Deaf Communication”. In Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta. pp 80-84.

Forster J., Stein D., Ormel E., Crasborn O., Ney H.. 2010. “Best Practice for Sign Language Data Collections Regarding the Needs of Data-Driven Recognition and Translation”. In Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta. pp 92-98.

Geraci C., Bayley R., Branchini C., Cardinaletti A., Cecchetto C., Donati C., Giudice S., Mereghetti E., Poletti F., Santoro M., Zucchi S. 2010. “Building a corpus for Italian Sign Language. Methodological issues and some preliminary results”. In Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta. pp 98-102.

Hanke T., König L., Wagner S., Matthes S., 2010. “DGS Corpus & Dicta-Sign: The Hamburg Studio Setup”. In Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta. pp 106-110.

Hansen B., 1975 “Varieties in Danish Sign Language, Sign Language Studies” 8 (1975) 249–256. J. Kyle, British Sign Language, Special Education 8 (1981) pp 19–23.

Herrero, A., 2004 “Escritura alfabética de la Lengua de Signos Española” Universidad de Alicante. Servicio de Publicaciones.

Herrero, A. 2009. “Gramática didáctica de la Lengua de Signos Española (LSE)”. Ed. SM, 2009.

Johnston T., 2008. “Corpus linguistics and signed languages: no lemmata, no corpus”. In Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages. Marrakech, Morocco.

Koehn P., Och F.J., Marcu. D.,2003. “Statistical Phrase-based translation”. In Proceedings of the Human Language Technology Conference 2003 (HLT-NAACL 2003), Edmonton, Canada, pp. 127-133.

Koehn, P., Hoang, H., “Factored Translation Models”. 2007. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague. Czech Republic. pp. 868–876.

Koehn, P,. 2010. “Statistical Machine Translation”. Cambridge University Press.

Lidell S., and Johnson, R., 1989 “American Sign Language: The phonological base”. Sign Language Studies, 64: pp 195-277.

Mariño J.B., Banchs R., Crego J.M., Gispert A., Lambert P., Fonollosa J.A., Costa-Jussà M., 2006. "N-gram-based Machine Translation", Computational Linguistics, Association for Computacional Linguistics. Vol. 32, nº 4, pp. 527-549.

Marshall, I., Sáfár, E. (2005) “Grammar Development for Sign Language Avatar-Based Synthesis”, In Proceedings of the 11th International Conference on Human Computer Interaction (HCII 2005), Las Vegas, USA.

Morrissey S., Way A., Stein D., Bungeroth J., and Ney H., 2007 “Towards a Hybrid Data-Driven MT System for Sign Languages. Machine Translation Summit (MT Summit)”, Copenhagen, Denmark. pp 329-335.

Morrissey, S. 2008. “Data-Driven Machine Translation for Sign Languages”. Thesis. Dublin City University, Dublin, Ireland.

Morrissey S., Somers H., Smith R., Gilchrist S., Dandapat S., 2010 “Building Sign Language Corpora for Use in Machine Translation”. In Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta. pp 172-178.

Notoya M., Suzuki S., Furukawa M., Umeda R., 1986 “Method and acquisition of sign language in profoundly deaf infants”, Japan Journal of Logopedics and Phoniatrics 27 (1986) pp 235–243.

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

Och J., Ney. H., 2003. “A systematic comparison of various alignment models”. Computational Linguistics, Vol. 29, No. 1 pp. 19-51.

Papineni K., Roukos, S., Ward, T., Zhu. WL., 2002 “BLEU: a method for automatic evaluation of machine translation”. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, USA, pp. 311-318.

Penn, C. Lewis, R. Greenstein A., 1984. ”Sign Language in South Africa”, South African Disorder of Communication 31 (1984), pp 6–11.

Pinedo, FJ., 2000. “Diccionario de Lengua de Signos Española”. Ed. CNSE Confederación de Personas Sordas Española.

Pizzuto, E., Rossini, P., & Russo, T. 2006. “Representing signed languages in written form: questions that need to be posed”. In C. Vettori (ed.). In Proceedings of the “Second Workshop on the Representation and Processing of Sign Languages”- LREC 2006 – 5th International Conference on Language Resources and Evaluation. Pisa, Italy: ILC-CNR, pp 1-6.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T., Henning, J., 1989. “Hamburg Notation System for Sign Languages – An introductory Guide”. International Studies on Sign Language and the Communication of the Deaf, Volume 5. Institute of German Sign Language and Communication of the Deaf, University of Hamburg. Germany.

San-Segundo R., Barra R., Córdoba R., D’Haro L.F., Fernández F., Ferreiros J., Lucas J.M., Macías-Guarasa J., Montero J.M., Pardo J.M, 2008. “Speech to Sign Language translation system for Spanish”. Speech Communication, Vol 50. pp 1009-1020.

San-Segundo, R., Pardo, J.M., Ferreiros, F., Sama, V., Barra-Chicote, R., Lucas, JM., Sánchez, D., García. A., "Spoken Spanish Generation from Sign Language" Interacting with Computers, Vol. 22, No 2, pp. 123-139.

Schembri. A., 2008 "British Sign Language Corpus Project: Open Access Archives and the Observer's" In Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages. Marrakech, Morocco.

Stein, D., Bungeroth, J. and Ney, H.: 2006 "Morpho-Syntax Based Statistical Methods for Sign Language Translation". In Proceedings of the 11th Annual conference of the European Association for Machine Translation, Oslo, Norway.

Stokoe, W., 1960 "Sign Language structure: an outline of the visual communication systems of the American deaf" Studies in Linguistics, Buffalo University, USA. Paper 8.

Stokoe, W,. Armstrong, C., Wilcox, D., Sherman E. 1995. "Gesture and the Nature of Language". Cambridge University Press.

Stolcke A., 2002. "SRILM – An Extensible Language Modelling Toolkit". In Proceedings of the International Conference on Spoken Language Processing, Denver, USA, vol. 2, pp. 901-904.

Sumita E., Akiba, Y., Doi, T., Finch, A., Imamura, K., Paul, M., Shimohata, M., Watanabe, T.,, 2003. "A Corpus-Centered Approach to Spoken Language Translation". In Proceedings of the Conference of the European Chapter of the Association For Computational Linguistics (EACL), Budapest, Hungary. pp 171-174.

Rodríguez MA., 1991. "Lenguaje de signos, PhD. Dissertation", Confederación Nacional de Sordos Españoles (CNSE) and Fundación ONCE, Madrid. Spain.

Vendrame M., Tiotto G., 2010. ATLAS Project: Forecast in Italian Sign Language and Annotation of Corpora. In Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta. pp 239-243.

Zens R., Och, FJ., Ney. H., 2002. "Phrase-Based Statistical Machine Translation". German Conference on Artificial Intelligence (KI 2002). Aachen, Germany, Springer, LNAI, pp. 18-32.

Wheatley, M., Annika P., 2010. "Sign Language in Europe". In Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. LREC. Valleta, Malta. pp 251-255.