A SPANISH SPEECH TO SIGN LANGUAGE TRANSLATION SYSTEM FOR ASSISTING DEAF-MUTE PEOPLE

R. San-Segundo, R. Barra, L.F. D'Haro, J.M. Montero, R. Córdoba, J. Ferreiros

Grupo de Tecnología del Habla. Universidad Politécnica de Madrid
{lapiz|barra|lfdharo|juancho|cordoba|jfl}@die.upm.es

10.21437/Interspeech.2006-420

ABSTRACT

This paper describes the first experiments of a speech to sign language translation system in a real domain. The developed system is focused on the sentences spoken by an officer when assisting people in applying for, or renewing the National Identification Document (NID) and the Passport. This system translates officer explanations into sign language for deaf-mute people. The translation system is composed by 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 gestures belonging to the sign language), and a 3D avatar animation module (for playing the gestures). The field experiments have reported a 27.2% GER (Gesture Error Rate) and a 0.62 BLEU (BiLingual Evaluation Understudy).

Index Terms: Machine Translation, Spanish Sign Language, Speech Translation, Gesture Animation

1. INTRODUCTION

The sign language presents a great variability depending on the country, even between different areas in the same country. Because of this, from 1960 sign language studies have appeared not only in USA [1][2][3] but also in Europe [4][5], Africa [6] and Japan [7]. In Spain, during the last 20 years, there have been several proposals for normalizing Spanish Sign Language, but none of them has been accepted by the deaf-mute people community. From their point of view, these proposals tend to constrain the sign language, limiting its flexibility. In 1991, MA. Rodríguez [8] carried out a detailed analysis of Spanish Sign Language (SSL). She showed the differences between the sign language used by deaf-mute people and the standardization proposals. This work is one of the main studies on SSL and the main reference in this work.

Spoken language translation has been and is being investigated in a number of join projects like C-Star, ATR. Vermobil, Eutrans, LC-Star, PF-Star and TC-Star. Apart from the project TC-Start (the last one), these projects addressed translation tasks with rather limited domains (like traveling and tourism) and medium sized vocabularies. The best performing translation systems are based on various types of statistical approaches [9], including example-based methods [10], finite-state transducers [11] and other data driven approaches. The progress achieved over the last 10 years is due to several factors like automatic error measures [12], efficient algorithms for training [13], context dependent models [10], efficient algorithms for generation [14], and more powerful computers and more parallel corpora.

The eSIGN (Essential Sign Language Information on Government Networks) European Project [15] constitutes one of the most important effort in developing tools for automatically generation of sign language contents. In eSIGN project, the main results has been a 3D avatar (VGuido) with enough flexibility to represent gestures from the sign language, and a visual environment for creating gesture animations in a easy way. The tools developed in eSIGN were oriented to translate web content into sign language. Sign language is the first language of many Deaf people, and their ability to understand written language may be poor in some cases. As such, it is very important for this group to have access to information in their first language, sign language. The result of the project is working on local Government websites in Germany, the Netherlands and United Kingdom.

In the recent years several groups have showed interest in machine translation for Sign Languages, developing several prototypes: example-based [16], rule-based [17], full sentence [18] or statistical [19] approaches. This paper includes the first experiments on one of the first speech to sign language translation systems, and the first one developed specifically for the Spanish Sign Language.

This paper is organized as follows: in section 2, an overview of the system is presented including a description of the task domain and the database. Section 3 describes the speech recognizer. In section 4, the natural language translation module is described. Section 6 shows the gesture playing module using a 3D avatar, and finally, section 7 summarizes the main conclusions of the work.

2. SYSTEM OVERVIEW

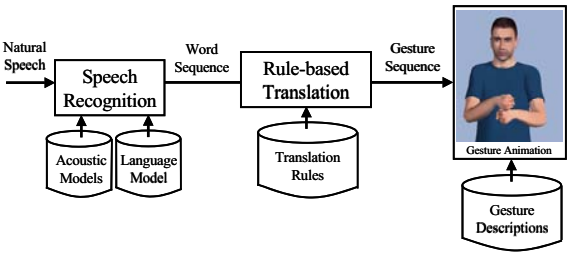

Figure 1.Spoken Language to Sign Language translation system.

Figure 1 shows the module diagram of the system. The first module, the speech recognizer, converts natural speech into a sequence of words (text) using acoustic and language models. The natural language translation module converts a

word sequence into a gesture sequence. This module consists of a rule-based translation strategy, where a set of translation rules (defined by an expert) guides the translation process. The gesture animation is performed by VGuido: the eSIGN 3D avatar developed in the eSIGN project [15], incorporated in the translation system as an ActiveX control. The gesture descriptions are generated through the eSIGN Editor.

2.1 Domain and database

The developed system is focused on a limited domain. This domain is composed by sentences spoken by an officer when assisting people in applying for, or renewing the National Identification Document (NID) and the Passport. In this context, a speech to sign language translation system is very useful because most of the officers do not know the sign language and they have problems when interacting to deaf-mute people. This system translates the officer explanations into sign language to provide a better service.

For developing the system, the most used phrases has been selected from normal dialogues between officers and users (135 phrases). These sentences contain more than 458 different words. These sentences have been translated by hand into Spanish Sign Language (SSL) generating more than 270 different gestures (Table 1 summaries the corpus statistics). As a result of this process, a parallel corpora is generated: word sequences and their corresponding gesture sequences. In this work, every gesture has been represented by a word written in capital letters. For example, the sentence "you have to pay 20 euros as document fee" is translated into "FUTURE YOU PAY TWENTY EURO DOC_FEE".

Table 1. Corpus statistics summary.

	Spanish	SSL
Sentences Pairs	135	
Number of words	1606	1470
Vocabulary	458	270

3. SPEECH RECOGNITION

The speech recognizer used is a state of the art speech recognition system developed at GTH-UPM [20]. It is a HMMs (Hidden Markov Models) based system with the following main characteristics:

- It is a Continuous Speech recognition system: it recognizes utterances formed by several words continuously spoken. In this application, the vocabulary size is 458 Spanish words.
- Speaker independency: The acoustic HMMs have been trained with a very big database, containing more than 20 hours of speech from 4000 speakers. The size of the database and the variability of the speakers provide the acoustic models with an important recognition power and robustness.
- The recognition system can generate one optimal word sequence (given the acoustic and language models), a solution expressed as a directed acyclic graph of words that may compile different alternatives, or even the N-best word sequences sorted by similarity to the spoken utterance. In this work, only the optimal word sequence is considered.
- The recognizer provides one confidence measure for each word recognized in the word sequence. The confidence measure is a value between 0.0 (lowest confidence) and 1.0 (highest confidence) [21].

The speech recognizer uses 5760 triphone HMMs for modeling all possible allophones and their context (acoustic modeling). The system also has 16 silence and noise HMMs for detecting acoustic effects (non speech events like background noise, speaker artifacts, filled pauses,…) that appear in spontaneous speech. It is important to detect and process these effects in order to avoid that these noises affect the recognition performance.

The second source of knowledge included in a speech recognizer (besides the acoustic model) is the language model. This model complements the acoustic knowledge with the information about the most probable sequences of words. In this system, the recognition module uses a bigram language model. The reason of using just bigrams is because there is a low number of sentences to train the model. The speech recognition results in this task are presented in Table 3.

Table 2. Final speech recognition results: Word Error Rate (WER), Insertions (INS), Deletions (DEL) and Substitutions (SUB).

WER (%)	INS (%)	DEL (%)	SUB (%)
9.6	**1.8**	**3.8**	**4.0**

4. NATURAL LANGUAGE TRANSLATION

In this approach, the natural language translation module has been implemented using a rule-based technique considering a bottom-up strategy. The relations between gestures and words are defined by hand employing an expert. In a bottom-up strategy, the translation analysis is performed starting from each word individually and extending the analysis to context words or already-formed gestures (generally named blocks). This extension is done to find specific combinations of words and/or gestures (blocks) that generate another gesture. Not all the blocks contribute (or with other wording, need to be present) to the formation of the final translation. 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 to one or several syntactic-pragmatic tags. After that, the translation module applies different rules that convert the tagged words into gestures by means of joining words or gestures (blocks) and defining new gestures. At the end of the process, the block sequence must correspond to the gesture sequence resulting from the translation process.

Considering the 4 situations reported in [22], it is possible to classify the rules in 4 types:

- One word corresponds to an specific gesture: In this case, one word is directly mapped onto a specific gesture. Some examples are the numbers (one, two, …) and some substantives: photograph, policeman,…

- Several words generate a unique gesture. Some examples are paraphrases like “police office” or complex names like “Community of Madrid”.
- In the third type, one word generates several gestures. This situation appears in many translation issues like verbs, general and specific nouns, lexical-visual paraphrases, complex signs,…For example: the verb “necesitarás (you will need)“ is translated into the gesture sequence “FUTURO TU NECESITAR (FUTURE YOU NEED)”: one word is translated into 3 gestures.
- The last kind of rules are those that generate several gestures from several words with certain relationships between them. For example: “partida de nacimiento (birth document)” is translated into “DOCUMENTO TU NACER (DOCUMENT YOU BE_BORN)”.

The final version of the rule base translation module contains 153 translation rules written by an expert. The translation module has been evaluated with 135 utterances containing 458 words and 270 gestures. This evaluation has been performed by computing the percentages of correct gestures, inserted gestures (compared to the reference), deleted gestures (compared to the reference) and substituted gestures (translation into a wrong gesture). In order to compute these percentages, the translated gesture sequence is compared to the reference with a dynamic programming algorithm or Levenshtein distance, which considers equal costs for any kind of error. From these percentages, it is possible to compute the Gesture Error Rate (GER) in a similar way WER is computed in a speech recognition system. In this evaluation, the BLEU (BiLingual Evaluation Understudy) measure is also reported. This measure is less strict compared to GER and it is very used in machine translation research [12]. The final results are presented in Table 3.

Table 3. Final translation results.

	BLEU	GER (%)	INS (%)	DEL (%)	SUB (%)
TEXT	**0.79**	**16.8**	**4.2**	**10.2**	**2.4**
SPEECH	**0.62**	**27.2**	**6.5**	**17.8**	**2.9**

This table shows results in two different situations: considering directly the utterance transcription (TEXT) or considering the speech recognition output (SPEECH). As it is shown, the GER is higher when using the speech recognition output instead of the transcribed sentence. The reason is the speech recognizer introduces recognition mistakes that produce more translation errors: the percentage of wrong gestures increases (GER) and the BLEU decreases.

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

- In Spanish, it is very common to omit the subject of a sentence, but in Sign Language it is mandatory. In order to deal with this characteristic, several rules has been implemented in order to verify if every verb has a subject and to include a subject if there is any verb without it. When applying these rules some errors are committed: e.g. a wrong subject is associated to a verb.
- One sentence can be translated into different gesture sequences. When one of the possibilities are not considered in the evaluation, some errors are reported by mistake. This situation appears when the passive form is omitted in several examples.
- In Sign Language, a verbal complement is represented beginning with a specific gesture: for example a time complement is introduced with the gesture WHEN, or a manner complement is introduced with the gesture HOW. There are several rules for detecting the type of complement, but sometimes it is very difficult to detect if there is a location complement or a time complement. Also, it is necessary to omit the specific gesture when the verbal complement is very short (i.e. composed by one word: “today”, “now”, “here”,…). This is another cause of error when the complement length is wrongly estimated.

Apart from the aspects commented above, there are not many order problems when translating Spanish into SSL because the former (word order) has made an important influence in the latter (gesture order).

The rules developed for this domain have been classified in 3 levels depending on their domain dependency. Around 52% are general rules in SSL, 18% can be easily adapted to a similar domain and 30% domain specific.

5. GESTURE ANIMATION WITH THE ESIGN AVATAR: VGUIDO

The gestures are represented by means of VGuido (the eSIGN 3D avatar) animations. An avatar animation consists of a temporal sequence of frames, each of which defines a static posture of the avatar at the appropriate moment. Each of these postures in turn can be defined by specifying the configuration of the avatar’s skeleton, together possibly with some morphs which define additional distortions to be applied to the avatar (Figure 2).

Figure 2. Example of VGuido animation

In order to make an avatar sign or gesture, pre-specified animation sequences must be sent to the avatar. A signed animation is generated synthetically from an input script in the SiGML notation. SiGML (Signing Gesture Markup Language) is an XML application which supports the definition of sign sequences. The signing system constructs human-like motion from scripted descriptions of signing motions. These signing motions belong to “Gestural-SiGML”, a subset of the full SiGML notation, which is based on the HamNoSys notation for

Sign Language transcription [23]. HamNoSys and other components of SiGML mix primitives for static gestalts (such as parts of the initial posture of a sign) with dynamics (such as movement directions) by intention. This flexibility allows the transcriber to focus on essential characteristics of the signs when describing a sign. This information, together with knowledge about common aspects of human motion as used in signing such as speed, size of movement, etc., is also used by the movement generation process to compute the avatar's movements from the scripted instructions.

6. CONCLUSIONS

This paper has presented the first experiments of a speech to sign language translation system for a real domain. This domain consists of the sentences spoken by an officer when assisting people in applying for, or renewing the National Identification Document (NID) and the Passport. The translation system implemented is composed by a speech recognizer, and natural language translator and a gesture animator using a 3D avatar.

In these experiments, the natural language translator module consists of a rule-based translation module reaching a 27.2% GER (Gesture Error Rate) and a 0.62 BLEU (BiLingual Evaluation Understudy).

The rule-based translation module has presented a very high percentage of deletions compared to the rest of errors. This is due to the rule-based strategy: a speech recognition error makes that some word patterns do not appear (for fitting the defined rules) and some gestures are not generated. These errors have three main causes: subject omission, different translation alternatives and verbal complement detection.

7. ACKNOWLEDGEMENTS

Authors want 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. This work has been supported by TINA (UPM-CAM. REF: R05/10922), ROBINT (DPI2004-07908-C02) and EDECAN (TIN2005-08660-C04). Authors also want to thank E. Ibáñez and A. Huerta their contributions.

8. REFERENCES

[1] Stokoe, W., (1960). “Sign Language structure: an outline of the visual communication systems of the American deaf”. Studies in Linguistics. Buffalo.

[2] Christopoulos, C. Bonvillian, J., (1985). “Sign Language”. J. of Communication Disorders, 18 1-20.

[3] Pyers J.E., (2006) “Indicating the body: Expression of body part terminology in American Sign Language”. Language Sciences. Available online 4 January 2006.

[4] Hansen, B., (1975). “Varieties in Danish Sign Language”. Sign Language Studies, 8: 249-256.

[5] Kyle, J., (1981). “British Sign Language”. Special Education, 8: 19-23. 1981.

[6] Penn, C., Lewis, R., and Greenstein, A., (1984). “Sign Language in South Africa”. South African Disorder of Communication, 31: 6-11. 1984.

[7] Notoya, M., Suzuki, S., Furukawa, M., and Umeda, R., (1986). “Method and Acquisition of Sign Language in Profoundly Deaf Infants”. Japan Journal of Logopedics and Phoniatrics, 27: 235-243. 1986.

[8] Rodríguez, M.A. (1991). “Lenguaje de signos” Phd Dissertation. Confederación Nacional de Sordos Españoles (CNSE) and ONCE. Madrid. Spain. 1991.

[9] 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.

[10] 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.

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

[12] Papineni K., S. Roukos, T. Ward, W.J. Zhu. (2002) “BLEU: a method for automatic evaluation of machine translation”. 40th Annual Meeting of the ACL, Philadelphia, PA, pp. 311-318. 2002.

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

[14] Koehn P., F.J. Och D. Marcu. (2003) “Statistical Phrase-based translation”. Human Language Technology Conference 2003 (HLT-NAACL 2003), Edmonton, Canada, pp. 127-133, May 2003.

[15] http://www.sign-lang.uni-hamburg.de/eSIGN/

[16] 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.

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

[18] 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.

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

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

[21] Ferreiros, J., R. San-Segundo, F. Fernández, L.F.D'Haro, V. Sama, R. Barra, P. Mellén. (2005) New Word-Level and Sentence-Level Confidence Scoring Using Graph Theory Calculus and its Evaluation on Speech Understanding. Interspeech. pp 3377-3380.

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

[23] Prillwitz, S., R. Leven, H. Zienert, T. Hanke, J. Henning, et-al. (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, 1989.