English-ASL Gloss Parallel Corpus 2012: ASLG-PC12

Achraf Othman, Mohamed Jemni
Research Laboratory LaTICE, University of Tunis
5, Av. Taha Hussein, B.P. 56, Bab Mnara, 1008 Tunis, Tunisia
E-mail: achraf.othman@ieee.org, mohamed.jemni@fst.rnu.tn

Abstract

A serious problem facing the community of researchers in the field of sign language is the absence of a large parallel corpus for signs language. The ASLG-PC12 project proposes a rule-based approach for building a big parallel corpus of English written texts and American Sign Language glosses. We present a novel algorithm that transforms an English part-of-speech sentence to an ASL gloss. This project was started in the beginning of 2011 as a part of the project WebSign, and it offers today a corpus containing more than one hundred million pairs of sentences between English and ASL glosses. It is available online for free to promote development and design of new algorithms and theories for American Sign Language processing, for example statistical machine translation and related fields. In this paper, we present tasks for generating ASL sentences from the Gutenberg Project corpus that contains only English written texts.

Keywords: American Sign Language, Parallel Corpora, Sign Language

1. Introduction

To develop an automatic translator or any other tool that requires a learning task for Sign Languages, the major problem is the collection of parallel data between text and Sign Language. A parallel corpus contains large and structured texts aligned between source and target languages. They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe. Since there is no standard and sufficient corpus for Sign Language (Morrissey & Way, 2007; Morrissey S. , 2008), to develop statistical machine translation that requires pre-treatment prior to the execution of the process of learning which needs an important volume of data.

For these reasons, we started to collect pairs of sentences between English and American Sign Language Gloss. And due to absence of data, especially in ASL and in other side there exists a huge data of English written text; we have developed a corpus based on a collaborative approach where experts can contribute in the collection and in correction of bilingual corpus and also in validation of the automatic translation. Experts are people that are authorized to validate translations and correct suggestions of translations. ASLG-PC12 project (Othman & Jemni, 2011) was started in 2010, as a part of the project WebSign (Jemni & El Ghoul, 2007) that carries on developing tools able to make information over the web accessible for deaf. The main goal of our project WebSign is to develop a Web-based interpreter of Sign Language (SL). This tool would enable people who do not know Sign Language to communicate with deaf individuals. Therefore, contribute in reducing the language barrier between deaf and hearing people. Our secondary objective is to distribute this tool on a non-profit basis to educators, students, users, and researchers, and to disseminate a call for contribution to support this project mainly in its exploitation step and to encourage its wide use by different communities.

In this paper, we review our experiences with constructing one such large annotated parallel corpus between English written text and American Sign Language Gloss –the ASLG-PC12 (Othman & Jemni, 2011), a corpus consisting of over one hundred million pairs of sentences.

The paper is organized as follow. Section 2 presents a brief description about American Sign Language Gloss. Section 3 presents methods and pre-processing tasks for collecting data from the Gutenberg Project (Lebert, 2008). We present two stages of pre-processing, in which each sentences had been extracted and tokenized. After, we present our method and algorithms for constructing the second part of the corpus in American Sign Language Gloss. Constructed texts were generated automatically by transformation rules and then corrected by human experts in ASL. We describe also the composition and the size of the corpus. Discussions and conclusion are drawn in section 5.

2. Background

Several projects, concerned with Sign Language, recorded or annotated their own corpora, but only few of them are suitable for automatic Sign Language translation due to the number of available data for learning and processing. The European Cultural Heritage Online organization (ECHO) published corpora for British Sign Language (Woll, Sutton-Spence, & Waters, 2004), Swedish Sign Language (Bergman & Mesch, 2004) and the Sign Language of the Netherlands (Crasborn, Kooij, Nonhebel, & Emmerik, 2004). All of the corpora include several stories signed by a single signer. The American Sign Language Linguistic Research group at Boston University published a corpus in American Sign Language (Athitsos, et al., 2010). TV broadcast news for the hearing impaired are another source of sign language recordings. Aachen University published a German Sign Language Corpus of the Domain Weather Report (Bungeroth, Stein, Dreuw, Zahedi, & Ney, 2006). In 2010, Sara et al., (Morrissey, Somers, Smith, Gilchrist, & Dandapat, 2010) published a multimedia corpus in Sign Language for machine Translation. In literature, we found many related projects

aiming to build corpus for Sign Language. Most of them are based on video recording and we cannot find textual data toward building translation memory. Textual data for Sign Language is not a simple written form, because signs can contain others information line eye gaze or facial expressions. So, for our corpus, we will use glosses to represent Sign Language. In the next section, we will present a brief description about glosses.

3. Glossing signs

Stokoe (Stokoe, 1960) proposed the first annotation system for describing Sign Language. Before, signs were thought of as unanalyzed wholes, with no internal structure. The Stokoe notation system is used for writing American Sign Language using graphical symbols. After, others notation systems appeared like HamNoSys (Prillwitz & Zienert, 1990) and SignWriting (Sutton & Gleaves, 1995). Furthermore, Glosses are used to write signs in textual form. Glossing means choosing an appropriate English word for signs in order to write them down. It is not a translating, but, it is similar to translating. A gloss of a signed story can be a series of English words, written in small capital letters that correspond to the signs in ASL story. Some basic conventions used for glossing are as follows:

- Signs are represented with small capital letters in English.
- Lexicalized finger-spelled words are written in small capital letters and preceded by the '#' symbol.
- Full finger-spelling is represented by dashes between small capital letters (for example, A-C-H-R-A-F).
- Non-manual signals and eye-gaze are represented on a line above the sign glosses.

In this work, we use glosses to represent Sign Language. In the next section, we will describe steps for building our corpus.

4. English-ASL Parallel Corpus

3.1 Problematic issues

As we say in the beginning, the main problem to process American Sign Language for statistical analysis like statistical machine translation is the absence of data (corpora or corpus), especially in Gloss format. By convention, the meaning of a sign is written correspondence to the language talking to avoid the complexity of understanding. For example, the phrase "Do you like learning sign language?" is glossed as "LEARN SIGN YOU LIKE?". Here, the word "you" is replaced by the gloss "YOU" and the word "learn-ing" is rated "LEARN". Our machine translate must generate, after learning step, the sentence in gloss of an English input.

3.2 Ascertainment and approach

Generally, in research on statistical analysis of sign language, the corpus is annotated video sequences. In our case, we only need a bilingual corpus, the source language is English and the language is American Sign Language glosses transcribed. In this study, we started from 880 words (English and ASL glosses) coupled with transformation rules. From these rules, we generated a bilingual corpus containing 800 million words. In this corpus, it is not interested in semantics or types of verbs used in sign language verbs such as "agreement" or "non-agreement". Figure 1 shows an example of transformation between written English text and its generated sentence in ASL. The input is "What did Bobby buy yesterday?" and the target sentence is "BOBBY BUY WHAT YESTERDAY?". In this example, we save the word "YESTERDAY" and we can found in some reference "PAST" which indicates the past tense and the action was made in the past. Also, for the symbol "?" it can be replaced by a facial animation with "WHAT". For us, we are based on lemmatization of words. We keep the maximum of information in the sentence toward developing more approaches in these corpora. Statistics of corpora are shown in Table 1. The number of sentences and tokens is huge and building ASL corpus takes more than one week.

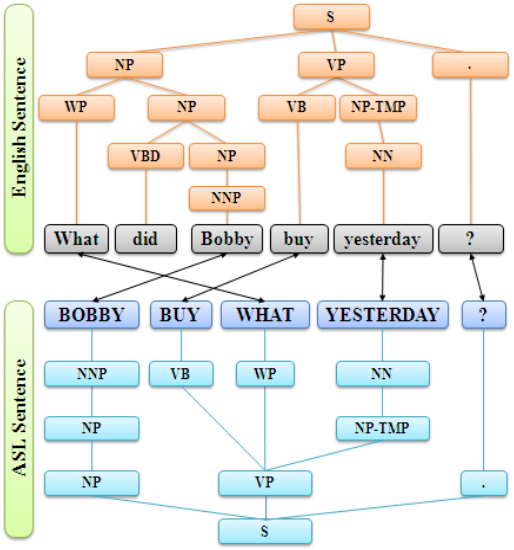

Figure 1: An example of transformation: English input 'What did Bobby buy yesterday?'

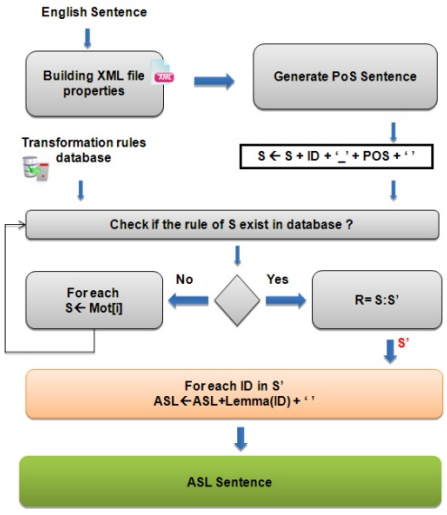

Figure 2: Steps for building ASL corpora

The input of the system is English sentences and the output is the ASL transcription in gloss. In table 2, only simple rules are shown, we can define complex rule starting from these simple rules. We can define a part-of-speech sentence for the two languages. According to figure 3, when we check if the rule of S exists in database, the algorithm will return true, in this case, we apply directly the transformation. Of course, all complex rules must be created by experts in ASL. Table 2 shows some transformation from English sentence to American Sign Language. We present the transformation rule made by an expert in linguistics.

	Corpus size English		Corpus size ASL Gloss	
	tokens	sentences	tokens	sentences
PART 1	280 M	13 M	280 M	13 M
PART 2	323 M	16 M	323 M	16 M
PART 3	549 M	27 M	549 M	27 M
PART 4	292 M	14 M	292 M	14 M
PART 5	150 M	7 M	150 M	7 M

Table 1. Size of the American Sign Language Gloss Parallel Corpus 2012 (ASLG-PC12)

English sentence: what is your name? **ASL sentence:** IX-PRO2 NAME, WHAT? **Transformation rule**: 1_VBP 2_PRP 3_JJ 4_. → 2_PRP 0_DESC- 3_JJ 4_.
English sentence: Are you deaf? **ASL sentence:** IX-PRO2 DESC-DEAF? **Transformation rule**: 1_VBP 2_PRP 3_DT 4_NN 5_. → 4_NN 2_PRP 5_.
English sentence: are you a student? **ASL sentence:** STUDENT IX-PRO2? **Transformation rule**: 1_VBP 2_PRP 3_DT 4_NN 5_. → 4_NN 2_PRP 5_.
English sentence: do you understand him? **ASL sentence:** IX-PRO2 UNDERSTAND IX-PRO3? **Transformation rule**: 1_VB 2_PRP 3_VB 4_PRP → 2_PRP 3_VB 4_PRP

Table 2. Example of full sentences transformation rules

In figure 2, we describe steps to transform an English sentence into American Sign Language gloss. The input of the system is the English sentence. Using CoreNLP tool, we generate an XML file containing morphological information about the sentence after tokenization task. Then, we build the part-of-speech sentence and thanks to the transformation rules database, we try to transform the input for each lemma. In some case, we can found that the part-of-speech sentence doesn't exist in the data-base, so, we transform each lemma. Transformation rule for lemma is presented in table 3. In the last step, we add an uppercase script to transform the output. The transformation rule is not a direct transformation for each lemma, it can an alignment of words and can ignore some English words like (the, in, a, an, etc.).

3.3 Transformations rules

Not all transformation rules used to transform English data were verified by experts in linguistics. We validate only 800 rules and transformation rules for lemma. We cannot validate all rules because there exist an infinite number of rules. For this reason, we developed an application that offer to experts to enter their rules from an English sentence, without coding. The application is just a simple user interface that contains lemma transformation rule, and the expert will compose lemma. After that, he save the result and rebuild the corpora. The built corpus is a made by a collaborative approach and validated by experts.

3.4 Collecting data from Gutenberg

Acquisition of a parallel corpus for the use in a statistical analysis typically takes several pre-processing steps. In our case, there isn't enough data between English texts and American Sign Language. We start collecting only English data from Gutenberg Project toward transform it to ASL gloss. Gutenberg Project (Lebert, 2008) offers over 38K free ebooks and more than 100K ebook through their partners. Collecting task is made in five steps:

Obtain the raw data (by crawling all files in the FTP directory).

- Extract only English texts, because there exist ebook in others languages than English like German, Spanish. We found also files containing ADN sequences.
- Break the text into sentences (sentence splitting task).
- Prepare the corpora (normalization, tokenization).

In the following, we will describe in detail the pre-processing steps to clean collected data.

3.5 Sentence splitting, tokenization, chunking and parsing

Sentence splitting and tokenization require specialized tools for English texts. One problem of sentence splitting is the ambiguity of the period "." as either an end of sentence marker, or as a marker for an abbreviation. For English, we semi-automatically created a list of known abbreviations that are typically followed by a period. Issues with tokenization include the English merging of words such as in "can't" (which we transform to "can not"), or the separation of possessive markers ("the man's" becomes "the man 's"). We use also an available tool for splitting called Splitta (Gillick, 2009). The models are trained from Wall Street Journal news combined with the Brown Corpus which is intended to be widely representative of written English. Error rates on test news data are near 0.25%. Also, we use CoreNLP tool (Toutanova & Manning, 2000; Klein & Manning, 2003). It is a set of natural language analysis tools which can take raw English language text input and give the base forms of words, their parts of speech.

3.6 Releases of the English-ASL Corpus

The initial release of this corpus consisted of data up to September 2011. The second release added data up to January 2012, increasing the size from just over 800 sentences to up to 800 million words in English. A forthcoming third release will include data up to early 2013 and will have better tokenization and more words in American Sign Language. For more details, please check the website (Othman & Jemni, 2011).

5. Discussions and conclusion

We described the construction of the English-American Sign Language corpus. We illustrate a novel method for transforming an English written text to American Sign Language gloss. This corpus will be useful for statistical analysis for ASL. We present the first corpus for ASL gloss that exceeds one hundred million of sentences available for all researches and linguistics. During the next phase of the ASLG-PC12 project, we expect to provide both a richer analysis of the existing corpus and others parallel corpus (like French Sign Language, Arabic Sign Language, etc.). This will be done by first enriching the rules through experts. Enrichment will be achieved by automatically transforming the current transformation rules database, and then validating the results by hand.

6. References

Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Thangali, A., et al. (2010, May 22-23). Large Lexicon Project: American Sign Language Video Corpus and Sign Language Indexing/Retrieval Algorithms. *" Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, LREC* .

Bergman, B., & Mesch, J. (2004). *ECHO data set for Swedish Sign Language (SSL).* Department of Linguistics, University of Stockholm.

Bungeroth, J., Stein, D., Dreuw, P., Zahedi, M., & Ney, H. (2006). A German Sign Language Corpus of the Domain. *Fifth International Conference on Language Resources and Evaluation*, (pp. 2000-2003). Genoa, Italy.

Crasborn, O., Kooij, E. v., Nonhebel, A., & Emmerik, W. (2004). *ECHO data set for Sign Language of the Netherlands (NGT).* Department of Linguistics, Radboud University Nijmegen.

Gillick, D. (2009). Sentence Boundary Detection and the Problem with the U.S. *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* , pp. 241-244.

Jemni, M., & El Ghoul, O. (2007). An avatar based approach for automatic interpretation of text to Sign language. *9th European Conference for the Advancement of the Assistive Technologies in Europe.* San Sebastian.

Klein, D., & Manning, C. D. (2003). Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics* , pp. 423-430.

Lebert, M. (2008). *Project Gutenberg (1971-2008).* University of Toronto & Project Gutenberg.

Morrissey, S. (2008). Assistive translation technology for deaf people: translating into and animating Irish sign language. *12th International Conference on Computers Helping People with Special Needs.* Linz.

Morrissey, S., & Way, A. (2007). Joining hands: developing a sign language machine translation system with and for the deaf community. *Conference and Workshop on Assistive Technologies for People with Vision and Hearing Impairments: Assistive Technology for All Ages.* Granada.

Morrissey, S., Somers, H., Smith, R., Gilchrist, S., & Dandapat, S. (2010, May). Building a Sign Language corpus for use in Machine Translation. *Proceedings of the 4th Workshop on Representation and Processing of Sign Languages: Corpora for Sign Language Technologies* , pp. 172-177.

Othman, A., & Jemni, M. (2011). *American Sign Language Gloss Parallel Corpus 2012 (ASLG-PC12).* Retrieved from http://www.achrafothman.net/aslsmt

Prillwitz, S., & Zienert, H. (1990). Hamburg notation system for sign language: Development of a sign writing with computer application. *nternational Studies on Sign Language and Communication of the Deaf* (pp. 355–379). Hamburg, Germany: Signum Press.

Stokoe, W. (1960). *Sign Language Structure: An Outline of the VisualCommunication Systems of the American Deaf.* Linstok Press, SilverSpring.

Sutton, V., & Gleaves, R. (1995). SignWriter— the world's first sign language processor. *Deaf Action Committee for SignWriting* .

Toutanova, K., & Manning, C. D. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. *the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* , pp. 63-70.

Woll, B., Sutton-Spence, R., & Waters, D. (2004). *ECHO data set for British Sign Language (BSL).* Department of Language and Communication Science, City University (London).