


Towards Developing Parallel Corpora for Portuguese and Portuguese Sign Languages

Student: Ziba Khani

Big data. Metodi statistici per la società della conoscenza, Master level,
Sapienza Università di Roma, Piazzale Aldo Moro, 5, 00185 Roma RM, Italy
khani.1799920@studenti.uniroma1.it

Mentors: Nuno Escudeiro 

Porto School of Engineering, Polytechnic Institute of Porto,
R. Dr. António Bernardino de Almeida 431, 4249-015 Porto, Portugal
nfe@isep.ipp.pt

Abstract. *Low-resource languages, including sign languages, are a challenge for machine translation research. Given the lack of large-scale parallel corpora, researchers must use small parallel corpora for training an automatic translation system. This article aims to address this problem by building artificial parallel corpora for Portuguese sign language in automatic translation systems. In this work, we obtained small parallel corpora of Portuguese text and Portuguese Sign Language gloss from the Metro of Porto. We used these corpora to learn grammar rules in translation between Portuguese text and Portuguese Sign language gloss. Applying obtained rules to our data, we generated artificial parallel corpora for Portuguese and Portuguese sign language gloss.*

Keywords. Sign language, low-resource languages, natural language processing, parallel corpora, machine translation

1 Introduction

Sign languages are natural languages that use the visual-manual modality to convey meaning through manual articulations and non-manual elements. They have their own grammar and lexicon and are not universal or mutually intelligible. Each sentence in sign language is composed of signs arranged according to a syntax governed by spatial and temporal logics, and each sign characterized by five parameters: configuration, orientation, location, movement, and facial expressions. To represent signs in text form, glossing is used, which involves capturing the essence of signs in written form.

This paper proposes a new approach to building parallel corpora for automatic translation systems by transforming a part of Portuguese speech sentences to Portuguese Sign Language (LGP) gloss. Related work, methodology, and conclusions are presented in Sections 2, 3, and 4 respectively.

2 Background

The lack of linguistic resources for Portuguese sign language poses a challenge for automatic processing such as machine translation and knowledge extraction [1]. Various studies have focused on collecting corpora for different sign languages, ranging from linguistic and humanistic to automatic translation. While some sign languages, such as American Sign Language (ASL), have rich linguistic resources [2],[3],[4], others, including Portuguese Sign Language, have limited annotated corpora, making it challenging to develop translation resources. Projects such as the Virtual Sign Translator [5] aim to address this issue by providing a translator between written Portuguese and Portuguese Sign Language. Other sign languages such as German, British, Spanish, French, and Irish also have their own annotated corpora – some projects focus on unique signs, and others focus on sentences and complete speeches. However, to the best of the authors' knowledge, there are no existing resources that can be used to build parallel corpora in text format for Portuguese sign language translation.

3 Parallel Corpora Collection

A parallel corpora is a collection of large and structured texts aligned between source and target languages, which is commonly used for statistical analysis and to validate linguistic rules within a specific domain. The process of acquiring a parallel corpora for statistical analysis usually involves several pre-processing steps. However, in our case, there is a lack of sufficient data available for Portuguese texts and Portuguese Sign Language.

The paper describes a methodology for creating a parallel corpora between Portuguese and Portuguese Sign Language (LGP) gloss using data from the Europarl dataset and a small parallel corpora from Porto's metro. The Europarl dataset [6] is a multilingual corpora that contains parallel text for 21 European languages, including Portuguese, extracted from the proceedings of the European Parliament. The researchers extracted Portuguese text from the Europarl dataset to create the parallel corpora.

The small parallel corpora from Porto's metro was translated by experts from the deaf community to ensure the correctness of sentences. The researchers used this corpora to extract rules for transforming Portuguese text into LGP gloss. They defined the translation problem as a series of sub-problems related to the order of words, word form changes, and lexical form changes due to gender in some languages. They obtained two sets of rules that can be seen as a mapping function, where the input is a representation of language, and the output shows the changes in input. Figure 1 shows the structure of the proposed system for rule extraction, which is composed of three predominant levels.

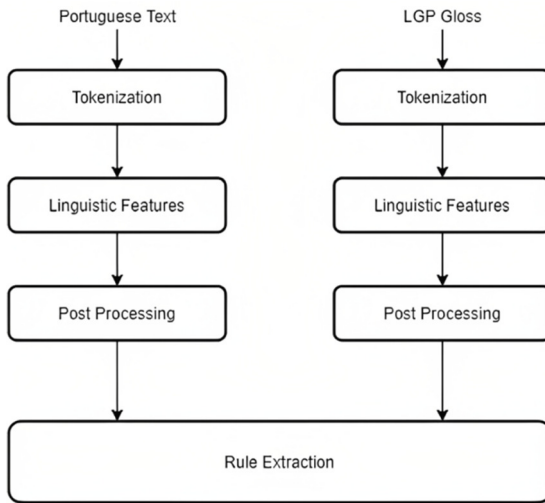


Figure 1. Extracting Grammar Rules.

The LGP gloss uses Portuguese words for each sign or phrase that can be labeled. The researchers describe the implementation phase, which involves pre-processing and lexical analysis, segmenting the text into sentences and words, and generating several analysis levels. The first pre-processing operation is called "tokenization," which precisely transforms the input string into tokens, as follows:

[“Como”; “abrir”; “uma”; “loja”; “no”; “Metro”; “?”]

Then, all characters are converted to lowercase, obtaining the following:

[“como”; “abrir”; “uma”; “loja”; “no”; “metro”; “?”]

From the lexical analysis, the researchers proceed to the grammatical analysis of each token. Syntactic analysis is the association of a grammatical category (noun, verb, adjective, adverb, proper noun, etc.) for each word of our input sentence.

In this step of the process, the SpaCy Part-Of-Speech Tagger is used to label grammatical features in the input text. The same process is applied to the LGP gloss to determine the grammatical features of each word in each sentence. The featurized sentences are then passed to a post-processing step that removes non-essential parts of the sentence, such as extra spaces

and symbols. The output of this process is a set of featurized sentences. The focus of the translation between Portuguese and LGP is on the first sub-problem of how the order of words changes. Rules are then extracted to indicate how the order of words changes, which are referred to as order-mapping. An example of the original sentence and the resulting order-mapping is provided:

Words	como	abrir	uma	loja	no	metro	?
Indices	0	1	2	3	4	5	6

Words	loja	metro	abrir	como	?
Indices	3	5	1	0	6

Given the example sentence and its corresponding LGP gloss, it is possible to translate the original sentence to LGP by analyzing how the order of indices changes. In this particular case, it was observed that the word at index 3 in the original sentence is located at index 0 in the LGP gloss, the word at index 5 in the original sentence is located at index 1 in the LGP gloss, and the word at index 2 in the original sentence does not exist in the LGP gloss. Based on this analysis, the researchers draw an order-mapping, which serves as a guide for translating sentences between Portuguese and LGP.

Indices of the original sentence	3	2	-1	0	-1	1	5
----------------------------------	---	---	----	---	----	---	---

The obtained rule for mapping can be used to reorder the words in the original sentence into an LGP sentence. However, the issue is that the rules are too specific to the given sentence. Instead, using features of the sentence, such as lemmas and postags, can create a more general map that can be applied to similar sentences. By using postags instead of words, the obtained map can be used for any new sentence with the same pattern (same postags).

The second sub-problem of the translation process involves transforming words between the original language and LGP. This is necessary because a word in the original text might not have a direct equivalent in LGP. To find pairs of similar words, the researchers used word embeddings to transform words into vectors of real numbers, and then measured their similarity using cosine similarity, a metric that measures the orientation between two vectors. This metric is independent of vector magnitude and can help find pairs of similar words in both languages.

In this step, the focus is on using cosine similarity to understand how a specific word is used differently across different corpora. The Glove word embedding model [7] is used to obtain vector representations for words in Portuguese language [8]. A cosine similarity of above 0.9 is used to determine word pairs. A dictionary of word mapping is created to indicate which words are translated to which words. Finally, an artificial parallel corpora is built using the rules extracted in the previous steps, and patterns are compared to generate LGP gloss for new Portuguese text. If any match is found, then it will be mapped to its counterpart mapping pattern, and the LGP gloss is generated in this way.

4 Conclusion

Sign Language Translation is a new research theme since it combines two complex scientific problems: translation and the transcription of Sign Languages. Studies on Sign Languages should include the linguistic, cognitive, and grammar aspects until creating the corpora, automatic translation, and real-time synthesis. Sign Languages are not universal, and in general, the studies are focused on one community of deaf and do not share the same syntactic structures, phonological, lexical, morphological, and semantic aspects. Despite existing tools for transcription and annotation, each presents drawbacks. However, for the textual annotation in gloss, we proposed an approach that uses the grammatical order of words to generate its counterpart LGP translation. We generated these texts automatically using rule-based approaches using the words' grammatical orders. The accuracy of these texts can be improved with getting access to the LGP dictionary and using linguistic guidance by experts in the future.

References

- [1] H. Y. Su and C. H. Wu, "Improving structural statistical machine translation for sign language with small corpora using thematic role templates as translation memory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1305–1315, 2009.
- [2] C. Neidle, A. Thangali, and S. Sclaroff, "Challenges in development of the American Sign Language Lexicon Video Dataset (ASLLVD) corpora," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, 2012.
- [3] T. J. Castelli, M. Betke, and C. Neidle, "Facial feature tracking and occlusion recovery in American sign language," in *Proceedings of the 2nd International Workshop on Pattern Recognition in Information Systems (PRIS)*, 2006.
- [4] C. Vogler and C. Neidle, "A new web interface to facilitate access to corpora: development of the ASLLRP data access interface," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, 2012.
- [5] P. Escudeiro, N. Escudeiro, R. M. Reis, M. Barbosa, J. Bidarra, A. B. Baltazar, and B. D. Gouveia, "Virtual sign translator," in *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology (ACE)*, 2013.
- [6] P. Koehn, "Europarl: A parallel corpora for statistical machine translation," in *Proceedings of the 10th Machine Translation Summit*, 2005.
- [7] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- [8] N. Hartmann, E. R. Fonseca, C. Shulby, M. V. Treviso, J. Rodrigues, and S. M. Aluísio, "Portuguese word embeddings: Evaluating on word analogies and natural language tasks," in *Proceedings of the 12th Brazilian Symposium in Information and Human Language Technology (STIL)*, 2017.

