Sign Language Translation Mobile Application and Open Communications Framework

Deliverable 4.3: First distributional intermediate representation based on embeddings - InterL-E

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101017255

Project Information
Project Number: 101017255
Project Title: SignON: Sign Language Translation Mobile Application and Open Communications Framework
Funding Scheme: H2020 ICT-57-2020
Project Start Date: January 1st 2021

Deliverable Information
Title: First distributional intermediate representation based on embeddings - InterL-E
Work Package: WP 4 – Transfer and InterLingual Representations
Lead beneficiary: University of the Basque Country (UPV/EHU)
Due Date: 30/04/2021
Revision Number: V1.0
Authors: Gorka Labaka, Santiago Egea, Euan McGill, Horacio Saggion
Dissemination Level: Public
Deliverable Type: Other

Document prepared by: UPF, UPV/EHU, TiU, KULeuven.

Overview: The purpose of this document is to describe the first version of the SignON project's embedding-based interlingua model (interL-E). This document presents the process of adapting mBART to the languages included in the project, as well as the results obtained by the fine-tuned model in the translation task for these languages.

Revision History

Version #	Implemented by	Revision Date	Description of changes
V0.1	Santiago Egea Gomez	18/04/2021	Initial draft after internal (WP4) review
V0.2	Gorka Labaka	25/04/2021	Partners' contributions

The SignON project has received funding from the European Union's Horizon 2020 Programme under Grant Agreement No. 101017255. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the SignON project or the European Commission. The European Commission is not liable for any use that may be made of the information contained therein.

Approval Procedure

Version #	Deliverable Name	Approved by	Institution	Approval Date
V0.2	D4.3	Aoife Brady	DCU	29/04/2021
V0.2	D4.3	Marcello Paolo Scipioni, Marco Giovanelli	FINCONS	29/04/2021
V0.1	D4.3	Vincent Vandeghinste	INT	20/04/2021
V0.2	D4.3	Gorka Labaka	UPV/EHU	26/04/2021
V0.2	D4.3	John O'Flaherty	MAC	27/04/2021
V0.2	D4.3	Irene Murtagh	TU Dublin	30/04/2021
V0.2	D4.3	Mathieu De Coster	UGent	26/04/2021
V0.2	D4.3	Jorn Rijckaert	VGTC	29/04/2021
V0.1	D4.3	Anthony Ventresque	NUID UCD	23/04/2021
V0.1	D4.3	Henk, Louis	RU	21/04/202x
V0.1	D4.3	Ineke Schuurman	KU Leuven	20/04/202x
V0.1	D4.3	Frankie Picron	EUD	22/04/2021
V0.1	D4.3	Dimitar Shterionov	TiU	18/04/2021

Acronyms

The following table provides definitions for acronyms and terms relevant to this document.

Acronym	Definition
GLUE	General Language Understanding Evaluation
KPI	Key Performance Indicator
MLM	Masked Language Models
MT	Machine translation
BLEU	Bilingual evaluation understudy
BERT	Bidirectional Encoder Representations from Transformers
TER	Translation Error Rate

Table of contents

1. Introduction

This document describes the advances achieved related to Work Package 4 (WP4) “Transfer and InterLingual Representations” in the scope of the project “SignON: Sign Language Translation Mobile Application And Open Communication Framework“. The main objective of this project is to provide an affordable application in order to remove communication barriers for deaf and hard of hearing people through automatic translation.

To facilitate the translation process, we will develop an InterLingual representation that will allow us to encode a message in one language (signed or verbal) and decode, i.e. translate, it into another language (signed or verbal).

Task 4.4 “Source or recognised text transformation from and to InterL” deals with both the encoding of a source language into an interlingual representation and the decoding, i.e. generation of language from the interlingual representation.

In this deliverable we describe the first interlingual representation - InterL-E. In particular, we present the model specification, our specific customisation, as well as the evaluation of such models on Spanish-English and Dutch-English MT tasks.

1.1. Background

Multilingual language models, such as mBART [5], XLM-R [7], M-BERT [2], and other variants have been widely adopted in many NLP tasks recently. These models use an efficient transformer-based architecture as a basic processing unit to create contextualised vectorial representations (on word- or sentence-level) for multiple languages in a unified embedding space. While such models are typically trained on large amounts of general-domain data and without a specific NLP task in mind, they can be specifically adapted or fine-tuned for a given objective - for example, text generation, question-answering, and other tasks beyond NLP [1] which deliver high performance for the given task. This process is known as transfer learning, whereby the knowledge gained from training a neural model for one task in turn is applied to another, related task. This approach assumes that using a previous task’s training will lead to greater efficiency in the time and resources required to train a new model, as well as increases in model performance.

Relevant models for the present task stem from BERT [2], which was the first such model that was released in 2018. BERT is pre-trained in a completely unsupervised way using large amounts of unlabelled data and taking into account the contexts of a word in either direction. Afterwards, the learned model is fine-tuned based on labelled data for the downstream task to complete. An advantage of this approach is its adaptability, and industry-leading GLUE[1][3] performance across NLP tasks.

However, other models have shown more promising results for MT tasks. For example, BART [4] has a similar architecture to BERT, but approaches pre-training differently: the unlabelled training data is masked and noised, before the model attempts to reconstruct the noisy utterances. This type of training has shown to be able to denoise, or decode, utterances into English - and achieves competitive results on MT (based on the BLEU metric).

More recently, mBART [5] was proposed to deal with multilingual modeling. mBART extends the BART pre-training phase employing a multilingual corpus including 25 languages, and can then be applied to both supervised and unsupervised MT tasks. mBART has shown performance improvements (based on the BLEU metric) over back-translation and other pre-trained model approaches. This is of particular interest to the present task. The BLEU metric is likely to be a key KPI (SignON WP1, Deliverable D1.13, due June 2021), and state-of-the-art performance on this benchmark makes this approach attractive for the project goals. Moreover, mBART performs substantially better with regard to its BLEU score on low and medium-resource language pairs. The verbal Irish language, and all sign languages involved in SignON are considered low-resource languages, and were not included in the original mBART model-building process. However, this process may be adapted as suggested by Tang et al. [13] by training for more iterations with the original data and new data containing new languages. The results look promising given the BLEU scores achieved in the original study. The application of mBART to WP4 is discussed in the following sections.

1.2. Transformer Layer

A transformer layer [8] is a module that takes a sequence of elements and outputs another sequence. In contrast to previous sequence-to-sequence architectures, the transformer does not employ recurrence, it is fully based on attention and it is also naturally bi-directional. The output of a transformer layer can

[1] General Language Understanding Evaluation score

be fed to another layer and, by stacking layers, an encoder-decoder model like the one depicted in Figure 1 can be created.

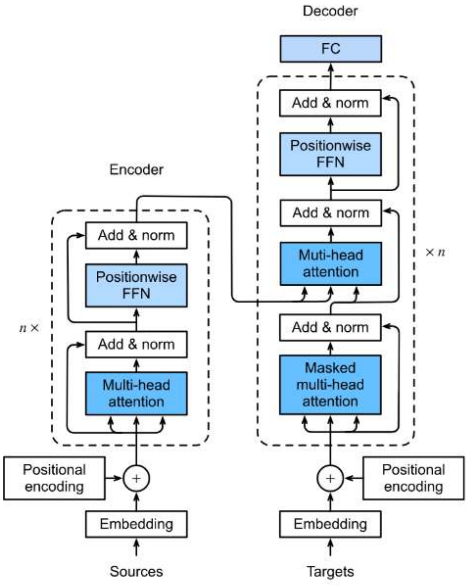

Figure 1. An encoder-decoder architecture based on Transformer layers [8]. A sentence is provided as input in the encoder; the output of the decoder is largely task-dependent, e.g., translation, text generation and so on.

The transformer-based module itself is mainly divided into two parts: the self-attention sub-module and the feed forward network. The former allows the model to learn a new representation for each word of the sequence based on how much attention the word puts on the other words (in both directions, left and right), without being conditioned on any external element. For instance, in the sentence “not all apples are red, they can also be green”, to discover what the “they” pronoun refers to, the model can look to the left to find the word “apples” and give particularly more attention to that word compared to the others. Moreover, this attention is not only computed once (single-headed attention), but multiple times (multi-headed attention).

Another important aspect of these layers is that not only do they require word embeddings at the first layer as their input, but also positional embeddings, as the module by itself does not understand the

order of words. These are created with sine and cosine functions to reflect the order of words within a sequence and are added to the word embeddings at the beginning.

1.3. Multilingual pretrained language models

Multilingual language models have been widely adopted in many NLP tasks. In this subsection we briefly present the best known models and the main differences between them:

- Multilingual BERT uses masked language models (MLM) to enable pre-trained deep bidirectional representations (Transformer). The training objective used is next sentence prediction. BERT constitutes a multi-layer bidirectional Transformer encoder, i.e. encoder-only. Although the multilingual BERT was trained on over 100 languages, it wasn't optimised for multi-linguality — most of the vocabulary is not shared between languages and therefore the shared knowledge is limited.
- XLM-R uses self-supervised training techniques where a model is trained in one language and then used with other languages without additional training data. It is trained on the MLM objective, as is BERT. XLM-R is trained on data in 100 languages and builds on the cross-lingual approach used with XLM [10] and RoBERTa [11] (i.e., pre-training with several monolingual sets and MLM, subword units to represent a shared vocabulary for all languages), but for a larger number of languages and with more training data (more than two terabytes of cleaned and filtered, publicly available CommonCrawl data [12]). The model has up to 550M parameters and is an encoder-only model.
- mBART is an encoder-decoder model. It is trained on CommonCrawl data for 25 languages. Noise is added to the input texts by masking phrases and permuting sentences, and a single Transformer model is learned to recover the texts, i.e. the BART objective. mBART is trained once for all languages, providing a set of parameters that can be fine-tuned for any of the language pairs in both supervised and unsupervised settings, without any task-specific or language-specific modifications or initialisation schemes [5].

The mBART model has shown significant improvements in terms of sentence-level machine translation (MT) for low- and medium-resource language pairs. To achieve such results the model needs to be fine-tuned as illustrated in Figure 2.

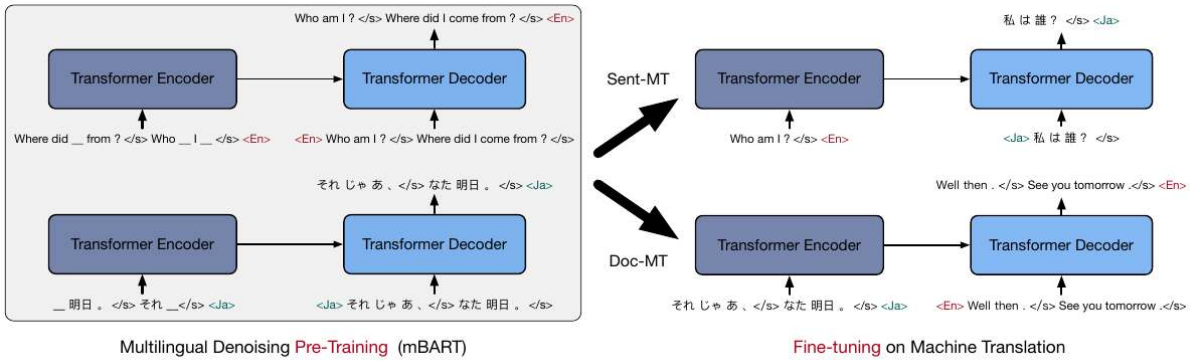

Figure 2: Fine-tuning of mBART for the task of MT [4]

2. First InterL-E Overview

The first interlingual representation, InterL-E, is based on mBART, an advanced Deep Learning model which uses transformer layers to manage long sentences. The success of transformers for modeling very long sentences is due to their capacity to learn complex sequence relations through the multi-headed attention.

Lately, NLP is largely dominated by pre-trained language models, such as BERT or XLM. This is especially the case in languages other than English, where multilingual models such as mBERT or XLM-R allow knowledge to be transferred from English to other languages with fewer resources. However, these models are not suitable for translation, as they only consist of an encoder capable of interpreting the input sentence. mBART, by contrast, is composed of both an encoder and a decoder, and has been trained on the generation of target text conditioned by some input, which is more suitable for translation. This is why we have decided to base our InterL-E on mBART.

Consequently, we have performed a fine-tuning to adapt the mBART model to the languages involved in the project. We selected mBART because it is pre-trained to manage multilingual translations, which fits with the objectives of the SignON project.

3. Fine-tuning for MT for SignON

Particularly, the mBART model is composed of 12 transformer layers for encoding and 12 transformer layers for decoding. Another fundamental feature of mBART is that it was pre-trained on a huge multilingual corpus and the learned weights are shared to take advantage of Transfer Learning. Thus, this model configuration is ideal for SignON.

Despite being multilingual, mBART does not use any parallel sentences during pre-training, where the model learns to reconstruct noisy sentences and where some words are masked and reordered. Therefore, the information from different languages is not aligned and the pre-trained model is not able to produce translations. In order to align the different languages and make mBART capable of producing quality translations, fine-tuning for translation is needed. This fine-tuning uses parallel sentences of the languages of interest to teach the model to translate between them, and thus align the information from the different languages.

However, due to the size of the vocabulary used in mBART (250,000 sub-words), fine-tuning the model is computationally expensive. So, when the fine-tuning is limited to a subset of the languages originally included in the model, it is possible to discard part of the vocabulary without losing performance. To do this, it is necessary to compile a representative corpus of the languages of interest and, after segmenting it using the original subword segmentation model provided with mBART, restrict the vocabulary to the subwords with more occurrences in such corpus. This results in a reduced vocabulary specialised in the languages/domains of interest, which can be used to limit both the segmentation model and the mBART model itself.

For the first version of InterL-E we have fine-tuned mBART for translating between three of the verbal languages included in the project and in mBART pre-training (English, Spanish and Dutch), while we will include Irish and signed languages in the next versions of InterL-E. We have used the English-Dutch (31M

segments) and English-Spanish (78M segments) ParaCrawl corpora [6] for fine-tuning, and reduced the vocabulary to the 40K tokens with the most occurrences in these corpora.

To combine the four translation directions in the same model we used iterative fine-tuning. We divided the corpus into batches of 30,000 parallel sentences and used these batches to tune the model in each of the translation directions in turn (Dutch->English, Spanish->English, English->Dutch and English->Spanish). We repeated the same process until 220 batches were used (6,600,000 parallel sentences in total). In this way, we managed to tune the model in all translation directions at the same time.

Nevertheless, the quality obtained in this way varies according to the order in which the language pairs are tuned. In our case, by tuning the last language pair (English->Spanish), some of the knowledge needed to translate the rest of the translation directions is forgotten. To measure this effect, we evaluated the translation performance of different fine-tuned models on a set extracted from the same Paracrawl corpus used in the training. Specifically, in addition to the final model resulting from the complete tuning process, one model per translation direction (saved after the last model update in that direction) was evaluated, as well as a model that combines them all by averaging the models.

3.1. Results

Table 1 shows the results achieved by the different models for the translation task between the languages for which fine-tuning has been performed. During fine-tuning, the corpora are iteratively used in the same order (Dutch->English, Spanish->English, English->Dutch and English->Spanish). The last pair of languages and direction used in the process is the English->Spanish (es->en). Thus, the model named “last model” and the language specific model for the es->en corpus are the same. In fact, they obtain the same results. Table 1 also shows an average model created taking four models and averaging their internal parameters. The models are taken from the last round of fine-tuning, for each of the corpora a model is saved.

	nl -> en		es -> en		en -> nl		en -> es	
	BLEU↑	TER↓	BLEU↑	TER↓	BLEU↑	TER↓	BLEU↑	TER↓
language specific	33.8	51.1	31.3	51.3	23.3	61.2	27.4	54.7
last model	31.7	54.5	23.7	64.2	19.0	69.4	**27.4**	**54.7**
average	**32.8**	**52.8**	**31.0**	**51.7**	**21.5**	**64.7**	25.7	55.6

Table 1. Results obtained by the different fine-tuned models for the translation task. Language-specific models are those in which the training ended with a specific pair of languages and a direction. The last model is the model obtained after finishing the training and the average model is created by averaging the weights of four models, one per language-pair and direction. Language-specific models are used as baselines. The best result using a single model is highlighted in bold.

The results show that using separate models for each translation direction gives the best results in all cases. On the other hand, using the model generated after the last fine-tuning step for all languages degrades the results in all cases except for the last direction being fine-tuned (English->Spanish). Finally, the model generated by averaging the language-specific models obtains more stable results. The averaged model outperforms the last model in all cases except English-Spanish, but still lags behind the separate models.

Taking into account that the aim of the present task is to obtain a common representation for all languages, maintaining separate representations per translation direction is not acceptable. We have therefore adopted the averaged model as the first version of the interL-E model.

3.2. Hardware Settings

We employed the high-performance computer resources offered by University of the Basque Country (UPV/EHU). Using a Linux server with four GPUs we executed the different experimental steps. The server is equipped with four Tesla V100 GPUs with 32 GB of memory.

4. Challenges, Limitations and Further Considerations:

At this stage we need to consider the following challenges:

- **Integration / fine-tuning to conversations (speech context):** Our use-cases are broadly related to conversations between individuals. Therefore, we will need to further adapt our InterL-E using conversational data, such as OpenSubtitles [9].
- **Integration / fine-tuning to recognised sign language:** This task depends on the input from the sign language recognition component from WP3. We will first rely on glosses (a tool to transcribe, in this case, sign language into text, capturing the facial and body grammar included in the signs - i.e. [14]) and fine-tune our model to recognise and translate from glosses into a verbal language, i.e., building natural sentences from glosses. One approach we are considering at this stage is a way of mapping subword units from the available vocabularies into sub-parts of glosses. This will make fine-tuning on glosses straightforward. Similarly, we will need to adapt or fine-tune our InterL-E to be able to output sequences in the Sign_A formalism that will be used to synthesise a 3D signer. That is, we would need to provide the model with parallel data such as the Spanish verbal language and the Spanish Sign_A formalism so that mBART understands how to translate from the source to the target. In a second stage, we will consider further integration of the InterL-E with the sign recognition component. Our initial approach considers sign recognition and translation as two separate components. However, the sign recognition is very likely to benefit from the language model that is inherent in our translation architecture. We will therefore investigate how the two might be integrated within an overarching approach. This topic will be investigated in two different ways. First of all, we will investigate how the output label probabilities of the sign recognition architecture might be explicitly reranked using information provided by the translation setup. Secondly, we will investigate whether we are able to integrate the output of the recognition setup directly as contextualized embeddings within our translation architecture.
- **Integration / fine-tuning to other supported languages**: At the moment we have developed our InterL-E to cover Spanish, English and Dutch. We will then look to integrate the Irish language too.

We should note also that integrating context as planned within WP4 is a challenge too. However this will be addressed in a later deliverable focused on the symbolic representation (InterL-S). Here context refers

to a broad range of additional information that can be used to further support and improve the transformation process. For example context can be the conversation history, speaker demographics (e.g. gender, age), conversation domain, etc. Additional linguistic information (syntactic, semantic, lexical diversity, etc.) will be considered.

5. References

[1] Bird, J. J. et al. (2020) Cross-Domain MLP and CNN Transfer Learning for Biological Signal Processing. *IEEE Access 8*. pp.54789-54801

[2] Devlin, J. et al. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 16pp.

[3] Wang, A. et al. (2018) Glue: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.* pp.353-355.

[4] Lewis, M. et al. (2019) BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. 10pp.

[5] Liu, Y. et al. (2020) Multilingual Denoising Pre-training for Neural Machine Translation. 17pp.

[6] Bañón M., Chen P., Haddow B., Heafield K., Hoang H., Esplà-Gomis M., Forcada M. L., Kamran A., Kirefu F., Koehn P., Ortiz Rojas S., Sempere L., Ramírez-Sánchez G., Sarrías E., Strelec M., Thompson B., Waites W., Wiggins D., Zaragoza J. (2020) ParaCrawl: Web-Scale Acquisition of Parallel Corpora. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4555–4567

[7] Conneau A., Khandelwal K., Goyal N. Chaudhary, Grave E., Ott M., Wenzek G., Zettlemoyer L., Guzmán F., Stoyanov V. (2020) Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.8440–8451

[8] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., L. Kaiser & Polosukhin I. (2017) Attention is All you Need. Advances in Neural Information Processing Systems 30 (NIPS).

[9] Lison P., Tiedemann J. (2016) OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)

[10] Conneau A., Lample G. (2019) Cross-lingual Language Model Pretraining. Advances in Neural Information Processing Systems (NIPS)

[11] Liu Y., Ott M., Goyal N., Du J., Joshi M., ChenD., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692

[12] Wenzek G., Lachaux M, Conneau A., Chaudhary V., Guzman F., Joulin A., Grave E. (2019) Ccnet: Extracting high quality monolingual datasets from web crawl data. arXiv preprint arXiv:1911.00359.

[13] Tang Y., Tran C., Li X., Chen P. J., Goyal N., Chaudhary V., Gu J., Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

[14] Porta, J., López-Collino, F. and Tejedor, J. (2014) A rule-based translation from written Spanish to Spanish Sign Language glosses. *Computer Speech & Language 28(3).* pp.788-811

