Reducing annotation effort for Cross-lingual Transfer Learning: The case of NLU for Basque

Maddalen López de Lacalle, Xabier Saralegi, and Inhar López

Elhuyar Foundation, Osinalde Industrialdea 3, 20170 Usurbil, Spain
{m.lopezdelacalle, x.saralegi, i.lopez}@elhuyar.eus

Abstract. Different cross-lingual transfer learning strategies have been proposed in the literature to address the two main tasks of natural language understanding (NLU), intent classification and slot filling. Most of these strategies are of the zero-shot type and offer much lower results than those obtained through training with data in the target language. This article analyzes the feasibility of an intermediate cross-language transfer learning strategy that only requires the manual annotation of a small number of examples in the target language. We have implemented different methods to select the most representative examples from the source dataset, based on clustering algorithms and explicit typological information. The selected examples are translated to the target language and annotated manually. Then, multilingual BERT models are fine-tuned to the both intent classification and slot filling tasks on the union of source language data and the manually annotated data in the target language. The approach has been evaluated on the *FMTOD* and *SNIPS* datasets, and compared to state-of-the-art zero-shot approaches. Evaluation focuses on an English to Basque transfer learning scenario which involves a less-resourced language with a different syntax and vocabulary from English. Results show that the proposed approach outperforms zero-shot approaches.

Keywords: Dialog systems · NLU · Neural Language Models · Less-Resourced Languages.

1 Introduction

Understanding the user's request is essential for a task-oriented dialog system. Typically a natural language understanding (NLU) module is responsible for doing so through two tasks: intent classification (extract the intent of the user utterance) and slot filling (extract the relevant slots of the intent). For instance, in the *"Cancel my doctor's appointment reminder for Friday"* utterance the user's intent is to cancel a reminder that was previously set, and the intent's arguments or slots are *"doctor's appointment"* and *"Friday"*.

Approaches based on neural network-based architectures [14, 1, 2] achieve best results on a variety of text classification tasks, intent classification included. As for the slot filling task, state-of-the-art approaches treat it as a sequence

labeling problem. Usually, bidirectional recurrent neural networks (BiLSTMs) are adopted with a subsequent Conditional Random Field (CRF) decoding layer [4, 8]. Some researchers propose to train jointly intent detection and slot filling [15, 9, 6, 3].

Elaborating utterance datasets with manual annotations of intent and slot filling is a very expensive process. Hence, transfer learning strategies that allow implementing models for languages different from the training dataset are of great interest, and even more so when the target language is a less-resourced language. There is some previous work on the matter at hand. Schuster et al. [12] analyze three different approaches for cross-lingual transfer learning, and conclude that given several hundred training examples in the target language, using cross-lingual pre-trained embeddings and their novel approach of using a multilingual machine translation encoder as contextual word representations outperforms translating the training data. López de Lacalle et al. [5] project the existing annotations in rich-resourced languages to the less-resourced language by means of Neural Machine Translation and posterior word alignments. Liu et al. [7] translate only a few task-related words selected from the scores computed by the attention layer of a trained English task-related model, in order to generate mixed-language sentences in the training data and learn the inter-lingual semantics across languages. Their approach outperforms [12] on the *FMTOD* dataset.

This paper studies a cross-lingual transfer learning approach that only requires the manual translation and annotation (intents and corresponding slots) in the target language of a small set of examples (utterances) from the source dataset. We assume that many of the examples included in the datasets of intent classification and slot filling tasks are variants of each other and that annotating manually the most prototypical ones in the target language is enough to significantly improve the cross-lingual transfer learning. Our hypothesis, based on other works [13, 16, 11], is that we are able to generalize the semantic structures of the examples included in the dataset by using a clustering process, and that we can select for each structure the most representative examples (the ones closest to the centroid of the cluster). These examples are manually translated into the target language and the information regarding intention and slots is also manually annotated. Then, we fine-tune multilingual BERT (Bidirectional Encoder Representations from Transformers) [2] models for intent classification and slot filling tasks using training data from the source language combined with the examples annotated manually in the target language. We pay special attention to the following aspects of our approach:

- How many examples should be annotated in the target language, and how should their semantic structures be generalized to select the most representative ones? To answer these questions we will analyze several cluster amounts, different selection criteria, and different measures of semantic distance.
- How is the transfer learning process affected by the low presence of the target language in the BERT multilingual model? To analyze this we compare a BERT multilingual model covering a large number of languages (mBERT

[2]) with another BERT multilingual model (mBERTeus [10]) limited to a smaller number of languages where the target language is better represented (Basque).

2 Proposed approach

Our approach to tackle intent classification and slot filling tasks is based on fine-tuning [2] multilingual BERT models. As a training dataset we use the union of the source language train data D_s and a small subset D_t of the data that has been translated into the target language and annotated manually (intents and slots). That subset corresponds to the most representative examples of the dataset in terms of their semantic structures. Our hypothesis is that including those few examples will notably improve the transfer learning process, enough to avoid the manual annotation of a large number of examples.

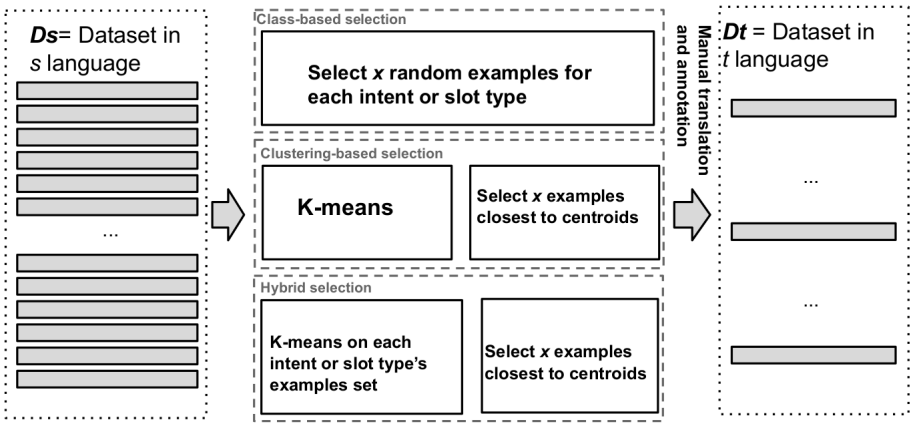

Fig. 1. Process of building target language training examples (D_t). We propose three approaches to select examples to translate from the source dataset D_s: a) *Class-based selection*, b) *K-means clustering-based*, c) *Hybrid selection*.

We analyzed three different approaches (see Figure 1) for selecting the most representative examples:

(a) *Class-based selection*: A simple way to build a representative sample is to select randomly a limited number of examples for each type of intention and slot from the source dataset D_s. Those examples are randomly selected from each type of intention or slot subset in the dataset.
(b) *K-means clustering-based selection*: We also implement a more sophisticated method based on clustering to identify the most representative examples. First, we cluster all the examples in the source dataset D_s using K-means.

Then, those closest to the centroids of the clusters are selected as prototypical examples. We analyze three different types of textual representation: a) Bag-of-words based on TFIDF, b) FastText, which is a dense representation based on static word embeddings, and c) SBERT, which is a dense representation based on contextual embeddings.

(c) *Hybrid selection*: The combination of both class-based and K-means based selections. The K-means selection is applied to the subsets of examples of each class. Instead of clustering all the examples without any restriction, each subset belonging to each type of intent or slot is clustered separately, in order to combine the strengths of both approaches.

When clustering with K-means the parameter k is set to match the number of examples selected and annotated by the class-based selection method and thus all approaches can be fairly compared.

Finally, we fine-tune pre-trained multilingual BERT language models on the intent classification and slot filling tasks independently using the union of D_s and D_t as training data. We have analyzed two multilingual pre-trained BERT language models containing a different representations of Basque: the official multilingual BERT model (henceforth mBERT) [2] which has been pre-trained on 104 different languages, and mBERTeus, a multilingual BERT model released by Otegi et al. [10] which has been pre-trained solely on Basque, Spanish and English. The presence of Basque in mBERTeus is higher than in mBERT, not only because its relative presence is larger (4 vs. 104 languages), but also because it includes a larger volume of Basque texts in absolute numbers. Specifically, mBERTeus is trained on a corpus containing 224.6 million tokens, of which 35 million come from the Wikipedia and the rest from online newspapers news articles.

In our experiments, the fine-tuning procedure used is the same as the one proposed in [2] which consists of feeding the output [CLS] token representation to an output layer for classification. In the case of the sequence labeler, the representations of output tokens are fed into a CRF layer.

3 Experiments and Results

The following two datasets have been chosen to carry out the experiments:

- ***FMTOD*** [12] contains manually generated and annotated utterances for three languages: English (*FMTODen* (43k)), Spanish, and Thai. They are grouped into three domains (alarm, reminder, and weather) and classified according to 12 types of intentions that include up to 11 types of slots. To evaluate our approaches we use the Basque version of the test of this dataset published by [5].
- ***SNIPS***[1] dataset considers 7 intent types (*AddToPlaylist*, *BookRestaurant*, *GetWeather*, *PlayMusic*, *RateBook*, *SearchCreativeWork* and *SearchScreen-*

[1] https://github.com/snipsco/nlu-benchmark/tree/master/2017-06-custom-intent-engines

Event) and 39 types of slots. Each intent type contains approximately 2k utterances for English in the train set. The train split contains 13,784 utterances. We manually translated the full test set (700 utterances) into Basque (*SNIPSeu*) and made it publicly available[2] .

All the evaluated systems are based on fine-tuning multilingual BERT models with English data jointly with a small set of examples translated into Basque and annotated manually. These examples (except for zero-shot strategies) have been selected following the strategies described below:

- **Zero_shot_direct**: Multilingual BERT model fine-tuned with source data. The models are fine-tuned for the specific tasks solely on English data. Therefore, it is not included any example translated into the target language in the training dataset.
- **Zero_shot_Liu** [7]: Model fine-tuned with English data, and using a seed list of English-Basque word pairs elaborated manually. This list includes most attended English words in English fine-tuning procedure, and it is used to build mixed-language data from English data. We evaluate this approach only on *FMTOD* dataset in order to reuse their attention-based selection of English words.
- **Random-selection (RS_{x})**: x examples, for manual translation and annotation, are selected randomly among the entire dataset. This will allow us to compare the other selection methods against a complete random example selection approach.
- **Class-based_sel_1**: A single example is randomly selected for each type of intent or slot, and then it is translated and annotated manually. Exactly, 12 and 7 examples have been selected from *FMTODen* and *SNIPS* for intent classification, respectively, and 11 and 39 examples from *FMTODen* and *SNIPS* for slot-filling, respectively.
- **Class-based_sel_2**: Equivalent to **Class-based_sel_1**, but selecting 4 random examples for each type of intent or slot. Exactly, 48 and 28 examples have been translated from *FMTODen* and *SNIPS* respectively for intent classification. For slot-filling, 44 examples from *FMTODen* were selected. This system was not evaluated for *SNIPS*, as its evaluation required to annotate a large number of examples.
- **{*Representation*}_{k}**: The entire dataset is grouped into k clusters. For each cluster a single example is selected, the one closest to the centroid, according to one of the following measures: the cosine similarity and TFIDF, FastText, or SBERT representations. In order to compare this method with the class-based and hybrid methods the number of clusters is set according to the amount of intention or slot types included in the dataset. For example, if the *FMTODen* includes 12 different types of intentions, 12 clusters will be generated. We also multiply the number of clusters by 4. Thus, in the previous case we will also group the dataset in 48 clusters.

[2] https://hizkuntzateknologiak.elhuyar.eus/assets/files/snipselh.tgz

– **Hybrid_{*Representation*}**: Class-based and clustering-based selection methods are combined so that each intent and slot type examples subsets are clustered separately. First, each subset of examples that belong to an intent or slot type is grouped into 4 clusters. Then, for each cluster the closest example to the centroid is selected according to the cosine similarity and TFIDF/FastText/SBERT representation. Exactly, 48 (4x12) and 28 (4x7) examples have been selected and translated from *FMTODen* and *SNIPS* respectively for intent classification. For slot filling, 44 (4x11) examples from *FMTODen* were translated. This system was not evaluated for *SNIPS*, as its evaluation required to translate and annotate manually a large number of examples.

BERT fine-tuning was done with a learning rate of 2e-5, and a batch size of 16, and all the experiments results shown in tables 1, 2, 3 and 4 present the averaged result of 5 randomly initialized runs. Slot filling was evaluated using official CoNLL evaluation script[3].

Method / Model	mBERT	mBERTeus	x
Zero_shot_direct	31.48	72.49	0
Zero_shot_Liu	56.61	77.55	-
RS_12	40.61	71.72	12
Class-based_sel_1	61.18	84.54	12
TFIDF_12	49.11	73.17	12
FastText_12	42.43	72.04	12
SBERT_12	48.42	74.90	12
RS_48	53.59	75.73	48
Class-based_sel_2	73.82	87.91	48
TFIDF_48	60.10	81.23	48
FastText_48	59.67	84.09	48
SBERT_48	64.64	76.98	48
Hybrid_TFIDF	80.62	92.49	48
Hybrid_FastText	78.23	93.25	48
Hybrid_SBERT	**82.94**	**94.05**	48

Table 1. Micro F1 results for Intent classification on *FMTOD* dataset. The last column x stands for the amount of examples manually translated and annotated into the target language.

Method / Model	mBERT	mBERTeus	x
Zero_shot_direct	22.62	74.33	0
Zero_shot_Liu	64.63	84.29	-
RS_11	37.35	79.87	11
Class-based_sel_1	61.08	83.66	11
TFIDF_11	42.69	78.5	11
FastText_11	54.26	80.85	11
SBERT_11	58.69	80.12	11
RS_44	**71.3**	83.21	44
Class-based_sel_2	67.55	84.07	44
TFIDF_44	66.33	84.34	44
FastText_44	70.21	85.12	44
SBERT_44	69.08	84.04	44
Hybrid_TFIDF	70.59	85.54	44
Hybrid_FastText	68.88	**86.06**	44
Hybrid_SBERT	69.16	84.46	44

Table 2. Micro F1 results for Slot filling on *FMTOD* dataset. The last column x stands for the amount of examples manually translated and annotated into the target language.

Table 1 and 2 show the results obtained on the *FMTOD* dataset for the intent classification and slot filling tasks respectively. The last column (x) indicates the number of examples selected from the source dataset that have been translated

[3] https://github.com/kyzhouhzau/BERT-NER/blob/master/conlleval.pl

and annotated. In general it is observed that the inclusion of target language examples in the training provides an improvement over zero-shot methods, especially over the *Zero_shot_direct* method. This improvement is more evident when mBERT is used as the base model. As expected, the more target language examples included in the training the better are the results.

For intent classification, the translation and annotation of 12 examples selected by the class-based approach is sufficient to overcome the zero-shots methods. However, for the slot filling task, more examples need to be translated.

Regarding the strategies to select more representative examples, in general, the hybrid method is the one that offers better results. The strategy of selection based only on classes is also competitive. And the strategy based on clustering only manages in some cases to overcome the latter in slot filling task. There are no significant differences in the use of TFIDF, FastText and SBERT. Note that in the case of intent classification, the way of selecting the examples is of great importance, since the random selection of 48 examples does not improve the zero-shot methods and a greater difference can also be observed among the results obtained by the three methods of example selection.

Method / Model	mBERT	mBERTeus	x
Zero_shot_direct	67.59	91.06	0
RS_7	67.28	90.63	7
Class-based_sel_1	73.46	91.05	7
TFIDF_7	77.83	91.49	7
FastText_7	78.91	91.17	7
SBERT_7	74.17	90.66	7
RS_28	75.6	93.4	28
Class-based_sel_2	83.77	**94.92**	28
TFIDF_28	80.37	91.37	28
FastText_28	81.8	94.06	28
SBERT_28	82.69	93.06	28
Hybrid_TFIDF	**84.54**	93.63	28
Hybrid_FastText	83.91	93.66	28
Hybrid_SBERT	82.11	92.59	28

Table 3. Micro F1 results for Intent classification on *SNIPS* dataset. The last column x stands for the amount of examples manually translated and annotated into the target language.

Method / Model	mBERT	mBERTeus	x
Zero_shot_direct	21.73	48.68	0
RS_39	39.26	55.29	39
Class-based_sel_1	37.48	55.72	39
TFIDF_39	40.65	56.55	39
FastText_39	**41.29**	**56.69**	39
SBERT_39	40.65	56.55	39

Table 4. Micro F1 results for Slot filling on *SNIPS* dataset[5]. The last column x stands for the amount of examples manually translated and annotated into the target language.

[5] Systems that required annotation of a larger number of examples were not evaluated for *SNIPS*. Exactly, 156 (39x4) examples would need to be annotated for each method. In total, 1248 (156x8) examples.

The results for *SNIPS* dataset are shown in the Table 3 and 4. For intent classification, using mBERT the hybrid approach provides the best improvement but with mBERTeus the class-based selection gets the best results. For slot filling, K-means based example selection does provide an improvement over the rest of the systems.

Between mBERT and mBERTeus it can be seen that mBERTeus provides better results, and that mBERT benefits most notably by including target language examples in training. mBERTeus is quite competitive for attempted classification, even with the zero-shot model. It seems that the greater presence of Basque data in mBERTeus benefits the transfer learning process.

4 Conclusions

This work shows that annotating very few examples (48 at most) in the target language and including them in the training set provides a significant improvement in the process of cross-lingual transfer learning for the intent classification and slot filling tasks. Moreover, this improvement is greater if the representation of the target language in the multilingual model is limited (mBERT).

On the other hand, we have verified that the set of examples to be annotated manually must be representative of the original corpus and that the proposed hybrid selection strategy based on combining clustering and intent and slot type information is effective.

5 Acknowledgements

This work has been partially funded by the Basque Government (DeepText project, Elkartek grant no. KK-2020/00088) and the Provincial Council of Gipuzkoa (NeuroLagun project).

References

1. Adhikari, A., Ram, A., Tang, R., Lin, J.: Rethinking complex neural network architectures for document classification. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4046–4051 (2019)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
3. Goo, C.W., Gao, G., Hsu, Y.K., Huo, C.L., Chen, T.C., Hsu, K.W., Chen, Y.N.: Slot-gated modeling for joint slot filling and intent prediction. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 753–757 (2018)

4. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
5. López de Lacalle, M., Saralegi, X., San Vicente, I.: Building a task-oriented dialog system for languages with no training data: the case for basque. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 2796–2802. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.340
6. Liu, B., Lane, I.: Attention-based recurrent neural network models for joint intent detection and slot filling. Interspeech 2016 pp. 685–689 (2016)
7. Liu, Z., Winata, G.I., Lin, Z., Xu, P., Fung, P.: Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. arXiv preprint arXiv:1911.09273 (2019)
8. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1064–1074 (2016)
9. Mesnil, G., He, X., Deng, L., Bengio, Y.: Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In: Interspeech. pp. 3771–3775 (2013)
10. Otegi, A., Agirre, A., Campos, J.A., Soroa, A., Agirre, E.: Conversational question answering in low resource scenarios: A dataset and case study for basque. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 436–442 (2020)
11. Qian, L., Zhou, G.: Clustering-based stratified seed sampling for semi-supervised relation classification. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 346–355 (2010)
12. Schuster, S., Gupta, S., Shah, R., Lewis, M.: Cross-lingual transfer learning for multilingual task oriented dialog. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 3795–3805 (Jun 2019). https://doi.org/10.18653/v1/N19-1380, https://www.aclweb.org/anthology/N19-1380
13. Tang, M., Luo, X., Roukos, S.: Active learning for statistical natural language parsing. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 120–127 (2002)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
15. Xu, P., Sarikaya, R.: Convolutional neural network based triangular crf for joint intent detection and slot filling. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. pp. 78–83. IEEE (2013)
16. Zhu, J., Wang, H., Yao, T., Tsou, B.K.: Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). pp. 1137–1144 (2008)