Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 137-147 recibido 31-03-2023 revisado 15-05-2023 aceptado 16-05-2023

Strategies for bilingual intent classification for small datasets scenarios

Estrategias de clasificación bilingüe de intenciones para escenarios con conjuntos de datos reducidos

Maddalen López de Lacalle, Xabier Saralegi, Aitzol Saizar, Gorka Urbizu, Ander Corral
Orai NLP Technologies
{m.lopezdelacalle, x.saralegi, a.saizar, g.urbizu, a.corral}@orai.eus

Abstract: This paper explores various approaches for implementing bilingual (Spanish and Basque) intent classifiers in cases where limited annotated data is available. Our study examines which fine-tuning strategy is more appropriate in such resource-limited scenarios: bilingual fine-tuning on a small number of manually annotated examples; a monolingual fine-tuning that relies on data augmentation via paraphrasing; or a combination of both. We explore two data augmentation strategies, one based on paraphrasing language models and the other based on back translation. Experiments are conducted on multiple pre-trained language models in order to evaluate the suitability of both monolingual and multilingual language models. The different approaches have been evaluated on two scenarios: i) a real use case over procedures associated with municipal sports services; and ii) a simulated scenario from the multi-domain Facebook Multilingual Task-Oriented dataset. Results show that data augmentation based on back translation is beneficial for monolingual classifiers that rely on pre-trained monolingual language models. Combining bilingual fine-tuning of the multilingual model with the data augmented by back translation outperforms the monolingual model-based approaches for Basque.
Keywords: Neural language models, dialog systems, less-resourced languages, intent classification, data augmentation.

Resumen: Este artículo explora varios enfoques para implementar clasificadores de intención bilingües (castellano y euskera) en casos en los que se dispone de un número limitado de datos anotados. Analizamos cuál es la estrategia de ajuste más adecuada en un contexto donde los recursos son escasos: ajuste bilingüe sobre un reducido número de ejemplos anotados manualmente; ajuste monolingüe basado en el aumento de datos mediante paráfrasis; o la combinación de ambos. Exploramos dos estrategias de aumento de datos, una basada en modelos lingüísticos de generación de paráfrasis y la otra en la traducción inversa. Además, los experimentos se realizan con múltiples modelos lingüísticos pre-entrenados para evaluar la idoneidad de los modelos lingüísticos monolingües y multilingües. Los distintos enfoques se han evaluado en dos escenarios: i) uno real, que corresponde a los trámites asociados a servicios deportivos municipales, y ii) otro simulado a partir del conjunto de datos multidominio Facebook Multilingual Task-Oriented Dataset. Los resultados muestran que para los clasificadores monolingües que se basan en modelos lingüísticos monolingües preentrenados, el aumento de datos basado en la traducción inversa es beneficioso. En el caso del euskera, la combinación del ajuste bilingüe del modelo multilingüe con los datos aumentados mediante la traducción inversa supera a los enfoques basados en modelos monolingües.
Palabras clave: Modelos de lenguaje neuronales, sistemas de diálogo, lenguas con menos recursos, clasificación de intenciones, aumento de datos.

ISSN 1135-5948 DOI 10.26342/2023-71-11

1 Introduction

The use of chatbots in real scenarios is gradually spreading. These are agents that interact with the user using natural language and whose purpose is to assist the user in specific tasks (task-oriented chatbots) or simply to maintain a conversation (chitchat). Task-oriented chatbots are the most useful from a commercial point of view as they serve to automate tasks that originally required the assistance of human operators.

One of the main NLU tasks involved in a task-oriented chatbot is intent classification. The chatbot is able to classify the intent implicit in the user's utterance, and thus determine the response action. Different approaches have been proposed in the literature to address this task, the most successful being approaches based on fine-tuning pre-trained neural language models. However, the implementation of this approach requires annotated collections. Since intents are dependent on the application scenario, annotated collections need to be built each time a system is to be implemented in a new scenario. Leaving aside the cross-lingual transfer setups, this annotation effort is doubled if the scenario is bilingual, as is the case in many countries or regions with more than one official language. This implies a significant manual effort, which is often not affordable.

This work focuses precisely on that kind of scenario: bilingual scenarios where there is limited manual effort available, sufficient only to annotate small-sized collections. We have determined a real geographical area, the Autonomous Community of the Basque Country, where Spanish and Basque are official languages, and we have established two scenarios, one real and the other simulated from a general collection. The real scenario corresponds to procedures associated with municipal sports services in the city of San Sebastian, and the simulated scenario has been designed based on the Facebook Multilingual Task Oriented Dataset (*FMTOD*) (Schuster et al., 2018) which is composed of annotated utterances for three task-oriented domains *alarm, reminder*, and *weather*.

The research questions addressed in this paper are as follows:

- RQ1: Which training strategy is best suited to implement bilingual intent classifiers when small annotated collections are available?
  - Bilingual training.
  - Data augmentation by paraphrasing.
  - Combination of both.
- RQ2: Is it more appropriate to train monolingual or multilingual models?
  - Is a zero-shot approach feasible?
- RQ3: Is data augmentation by paraphrasing helpful?
  - By means of paraphrase models?
  - By means of back translation?

From here on the paper is structured as follows. Section 2 reviews related works. Section 3 describes the data used in this work. Next, in Section 4 we detail the experimental setup. Obtained results are described in sections 5, 6, and 7. Finally, Section 8 draws conclusions on the experiments carried out.

2 Related works

State-of-the-art approaches address the intent classification task as a document classification problem and best results are usually obtained with approaches based on deep neural networks (Vaswani et al., 2017; Adhikari et al., 2019; Devlin et al., 2019).

Facing the problem of missing labelled (even unlabelled) data is a well-known problem for anyone involved in an NLP project. The process of preparing datasets for training systems is essential for machine learning but at the same time, it requires a lot of time and manual effort. The less data we have, the less data to train, and the less likely we are to get accurate predictions for data that our model has not yet seen. Therefore, different ways of paraphrasing have been studied to automatically expand manually created initial datasets in order to generate more training data by means of Data Augmentation (DA) techniques. (Wei and Zou, 2019) base their approach on simple word replacements using knowledge bases like WordNet to augment training samples and improve classification performance.

Many DA approaches are based on using pre-trained language models for sentence paraphrase generation. (Kumar, Choudhary, and Cho, 2020) compared three types of pre-trained transformers models, auto-regressive models (GPT-2)(Radford et al., 2019), autoencoder models (BERT)(Devlin et al., 2019), and seq2seq models (BART)(Lewis et al., 2019), for conditional DA. The Seq2Seq model outperformed the rest of the models and showed that by simply prepending the class labels to text sequences they effectively condition the models to generate new examples. Anaby-Tavor et al. (2020) also used GPT-2 for augmenting the data for text classification tasks. They fine-tune the language generator model with very few examples per class. Then, given a specific class label as input, the fine-tuned language generator is capable of generating new sentences for the class which are then filtered using a classifier trained on the original dataset.

Other approaches to increase the volume of training data are based on neural machine translation (NMT) (Sokolov and Filimonov, 2020; Goyal and Durrett, 2020; Mallinson, Sennrich, and Lapata, 2017; Federmann, Elachqar, and Quirk, 2019). Sokolov and Filimonov (2020) presented an automatic natural language generator (NLG) system for paraphrasing inspired on MT encoder-decoder deep RNN and achieved significant improvements on NLU tasks such as intent classification when training the models on the data augmented with their paraphrases. Goyal and Durrett (2020) presented a method that provides explicit control over the syntax of the generated paraphrases for which they first encode the syntax tree of the input sentence and then use this representation to feed the decoder to generate possible reorderings. Mallinson, Sennrich, and Lapata (2017) introduced PARANET, a model for paraphrase creation based exclusively on NMT (RNN architecture), and showed how the bilingual pivoting method can be implemented with NMT. Federmann, Elachqar, and Quirk (2019) also address paraphrasing through pivot translation, and showed that NMT techniques, especially applying pivoting through related languages, provide a relatively robust source of paraphrases with a level of diversity comparable to that of expert human paraphrases. As an alternative, grammar-based approaches provide a different option to paraphrasing that allow language generation from a formal description of its semantics, for example by means of the Grammatical Framework (Ranta, 2004).

Jolly et al. (2020) propose a DA approach based on an interpretation-to-text paraphrase model combined with shuffling-based sampling techniques that enable generating training data for new features even with limited seed data. The model maps a unique representation of a set of paraphrases (defined as the shared interpretation) to all its possible realizations. During inference, the model is conditioned on a specific interpretation and generates a distribution of possible realizations or paraphrasis for the new intents.

Recently, prompt-based approaches have been studied to generate synthetic data using GTP-3-like generative models for classifier training (Wang et al., 2022; Meyer et al., 2022) in scenarios with very little training data. Wang et al. (2022) proposed a soft prompt-based data augmentation model demonstrating that synthetic data produced by their model can improve performance on low-resource NLU tasks, including sentence classification and sequence labeling tasks, where only a few labeled data are available. Meyer et al. (2022) conclude that although classifiers trained on a small amount of manual data perform better than classifiers trained on synthetic data generated via prompting general purpose models, they show that in situations where only little data and resources are available, the cost-benefit trade-off of using this kind of synthetic data may be beneficial.

3 Datasets

Application scenarios of the chatbots will determine, among other things, the language and the set of intentions that need to be considered when interpreting users' utterances. Furthermore, in the chatbot development phase, one aspect conditioning the development strategy is the number of training examples available. This number is determined by the effort available to tackle this task in the development process. The less effort available, the fewer training examples will be generated.

The scenarios we want to study in this work have the following common features: the capability of a bilingual interaction (Spanish and Basque) with the user, and the availability of a reduced set of training examples. We have focused on two specific scenarios:

1. A chatbot for the municipal sports ser-

vice of the San Sebastian City Council: This is a chatbot that responds in Spanish and Basque, and supports the detection of the intentions corresponding to the operations related to the municipal sports services. In this case, data scarcity is real, since the few manual examples were created for the purpose of this work.

2. A multi-domain chatbot: This is a chatbot that works in Spanish and Basque and supports the detection of intentions in the three domains (*alarm, reminder*, and *weather*) included in the Facebook Multilingual Task Oriented Dataset (*FMTOD*) (Schuster et al., 2018). In this case, data scarcity is simulated. Although the FMTOD dataset contains many more examples, only a few examples were chosen to simulate a scarce-data scenario similar to the other one.

The training dataset for the chatbot of the municipal sports services (SportServ[1]) was created in collaboration with the service's technicians. They defined the various fields (*sports card, courses, and sports facilities*) as well as the intentions to be taken into account in each of these fields. The municipal sports service groups together the various procedures (e.g. to register for a new card, register for a course or reserve a facility) into the fields mentioned above. Hence, 12 types of intentions related to the *sports card* were specified (e.g. changing personal data on the sports card, canceling beneficiaries of the card, or presenting paperwork proving to be a student or unemployed). Similarly, another 11 types of intentions were defined for the *courses* field (e.g canceling or registering for a course, editing course registration data, paying registrations). Finally, for the *sports facilities* field, 2 intentions were set: to reserve a facility and to cancel the previous reservation of a facility. Subsequently, native speakers of Basque and Spanish independently generated in each language about 10 examples of utterances for each intention. As shown in Table 1 roughly 250 seed examples were manually generated for each language (about 10 examples per intent class and language) for the municipal sports services chatbot.

An example for each field in the SportServ dataset is shown in Table 2. The first utterance expresses the intention to obtain a new sports card (*"I want to join the sports center*", in English), the second the intention to unsubscribe from a course (*"I don't want to continue in the paddle tennis course"*, in English) and the third the intention to reserve a facility (*"I want to reserve a time to go to the pool"*, in English).

For the simulated case, splits of the same size have been randomly generated from the FMTOD dataset (see Table 1), specifically, we make use of the Basque (FMTODeu) and Spanish (FMTODes) versions of the FMTOD dataset published by (López de Lacalle, Saralegi, and San Vicente, 2020). This dataset includes 3 domains (*alarm, reminder, and weather*) and is classified according to a total of 12 types of intentions. To simulate a reduced training set, about 21 examples for each intention were randomly selected from the original dataset.

Dataset	**Train**	**Dev**	**Test**
SportServeu	133	50	77
SportServes	124	50	74
FMTODeu	133	50	77
FMTODes	124	50	74

Table 1: Statistics of manual examples datasets used in the experiments.

An example for each field in the SportServ dataset is shown in Table 3. The first utterance expresses the intention to modify an alarm ("*Delay alarm for 30 minutes*", in English), the second the intention to cancel a reminder (*"Cancel reminders for today"*, in English) and the third the intention to ask for a weather forecast (*"Will it rain at night?"*, in English).

4 Experiments

BERT can be easily fine-tuned on downstream tasks (Devlin et al., 2019), by training for a few epochs on the data of the specific task. In particular, we fine-tune multiple BERT-based pre-trained language models in the task of classifying intentions. In our experiments we use the same fine-tuning strategy as the original paper, feeding the output [CLS] token representation to an output layer for classification.

[1] https://storage.googleapis.com/orai-nlp/datasets/chatbots/dkorai.tgz

Field	Intent Label	Utterance
Sports card	kt_nueva_inscripcion	*Kiroldegian apuntatu nahi dut*
Courses	cursillo_baja	*Ez dut nahi padeleko ikastaroan jarraitzea*
Sports facilities	inst_nueva_reserva	*Igerilekura joateko ordua erreserbatu nahi dut*

Table 2: Examples for different intent classes belonging to the various fields on SportServ dataset.

Domain	Intent Label	Utterance
Alarm	modify_alarm	*Atzeratu alarma 30 minutu*
Reminder	cancel_reminder	*Ezeztatu gaurko gogorarazleak*
Weather	find	*Gauean euria egingo du?*

Table 3: Examples for different intent classes belonging to the various domains on FMTOD dataset.

In Section 5 we base our experiments on both monolingual and multilingual BERT models to study whether it is preferable to fine-tune models that can deal with a single language or those that can handle multiple languages when only a reduced set of training examples is available. In Section 6 the experimentation is carried on Basque and Spanish monolingual models in order to determine whether the data augmentation by paraphrasing is helpful in scenarios of limited labeled data. Finally, in Section 7 a multilingual BERT model is used in the experiments for answering whether combining data augmentation on a bilingual fine-tuning process can be beneficial or not.

The fine-tuning process on small datasets can be unstable (Mosbach, Andriushchenko, and Klakow, 2020; Zhang et al., 2020; Dodge et al., 2020) and lead to unfair comparisons between different approaches. Therefore, optimizing the hyperparameters of models can be crucial in order to achieve the best performance for each combination of methods and datasets.

We use the *Population Based Training* (PBT) algorithm (Jaderberg et al., 2017) to optimize a set of hyperparameters for the evaluated systems. Specifically, the search space includes the most sensitive BERT hyperparameters: *learning-rate, training batch-size, number of training epochs, weight decay,* and *random seed.* For each system, by means of PBT algorithm, we sample 100 different configurations of hyperparameters values from a given distribution to explore the search space[2].

Finally, to prevent the selection of hyperparameters that only work well on the training data, but do not generalize well to the test data, we carry on a three-fold cross-validation process. The dataset is split three times into equally sized train/dev/test subsets. The model optimization (100 runs) and evaluation procedure is repeated three times for each system, with a different subset serving as training, validation, and test sample each time. The best performance of the three test sets is taken as the final score.

5 Multilingual vs. Monolingual models

In our experiments we compare the following monolingual and multilingual models:

- **Berteus**: it is a monolingual pre-trained BERT model released by (Agerri et al., 2020) which has been pre-trained solely on Basque. Specifically, Berteus is trained on a Basque corpus comprising Basque crawled news articles from online newspapers and Basque Wikipedia. The training corpus contains 224.6 million tokens, of which 35 million come from Wikipedia.
- **BETO**: it is a monolingual pre-trained model released by (Cañete et al., 2020) which has been pre-trained solely on Spanish. The training corpus contains

[2] *learning-rate*: a uniform value from 1e-5 to 5e5; *batch-size*: a value from 4, 8, 16 and 32; *epochs*: a value from 10 to 35; *weight decay*: a uniform value from 0.0 to 0.3 and *seed*: a value from 1 to 5.

about 3 billion words and the texts were collected from different sources including Wikipedia, United Nations, and Government journals, TED Talks, Subtitles, News Stories, among others.

- **mBERTeus**: it is a multilingual pretrained BERT model released by (Otegi et al., 2020) limited to Basque, Spanish and English. This multilingual model fits our needs better than the official multilingual BERT model (Devlin et al., 2019) since the presence of Basque in mBERTeus is higher than in mBERT, not only because its relative presence is larger (3 vs. 104 languages), but also because it includes a larger volume of Basque texts in absolute numbers.

Table 4 and Table 5 show the results for the intent classification task on the SportServ and FMTOD datasets, respectively. All the evaluated systems are based on fine-tuning pre-trained BERT models with a small set of examples annotated manually (see amounts in Table 1).

- **Zero_shot_direct(ZS)**: Multilingual BERT model fine-tuned with data in a language other than the target language. Consequently, we test a classifier fine-tuned with Spanish examples for classifying Basque utterances ($train_{es} \rightarrow test_{eu}$) and vice versa.
- **Monolingual(Mono)**: Monolingual (Mono_1) and multilingual (Mono_2) BERT model fine-tuned exclusively on target language data. We fine-tuned a monolingual pre-trained model for each language (Berteus or BETO for Basque and Spanish, respectively) and the multilingual BERT model (mBerteus).
- **Multilingual(Multi)**: Multilingual BERT model fine-tuned on manual examples on both languages.

For the dataset related to municipal sports services, the system with the best results depends on the target language (see Table 4). For the Basque language, the monolingual Berteus model fine-tuned only with Basque examples obtains the best results and the addition of Spanish data in the multilingual training of mBerteus does not manage to overcome it.

System	Model	Train	Test	micro F1
ZS	mBERTeus	es	eu	24.68
Mono_1	Berteus	eu	eu	**75.75**
Mono_2	mBERTeus	eu	eu	68.83
Multi	mBERTeus	eu+es	eu	71.43
ZS	mBERTeus	eu	es	36.00
Mono_1	BETO	es	es	68.67
Mono_2	mBERTeus	es	es	65.33
Multi	mBERTeus	eu+es	es	**74.67**

Table 4: Micro F1 results for the systems on SportServ dataset.

System	Model	Train	Test	micro F1
ZS	mBERTeus	es	eu	46.75
Mono_1	Berteus	eu	eu	86.15
Mono_2	mBERTeus	eu	eu	91.34
Multi	mBERTeus	eu+es	eu	**92.21**
ZS	mBERTeus	eu	es	66.22
Mono_1	BETO	es	es	90.54
Mono_2	mBERTeus	es	es	87.39
Multi	mBERTeus	eu+es	es	**93.72**

Table 5: Micro F1 results for the systems without data augmentation on FMTOD dataset.

For the FMTOD dataset, fine-tuning multilingual BERT models with Spanish and Basque data jointly surpasses all other systems (see Table 5).

In both datasets it can be appreciated that the zero-shot strategy is not feasible. Moreover, between the two monolingual approaches, the one based on the monolingual model outperforms the multilingual model, except for Basque FMTOD dataset.

6 *Data augmentation based on paraphrasis*

In this section, we apply two different strategies for automatically augmenting the training data by means of paraphrasing. In order to study if data augmentation by paraphrasing is helpful, we first paraphrase a few manual examples and then incorporate them into the fine-tuning process jointly with the manual examples.

Next, we introduce the two strategies for augmenting the data based on paraphrasing.

Back translation: This approach on using back translation by means of NMT models, in order to generate paraphrases. That is, we first translate the sentence into a different language from the source sentence and then translate it back into the source language. The result is a sentence equivalent to the original but with a different vocabulary and grammar. For example, starting from the sentence *"quiero eliminar mi preinscripción"* corresponding to the intent class related to canceling the registration of a course, we first translate it into Basque (*"nire aurreinskripzioa ezabatu nahi dut"*), and then we translate it back from Basque into Spanish and select different translation hypotheses returned by the neural translation system. The Spanish to Basque MT system used for our experiments is based on the default Base Transformer architecture (Vaswani et al., 2017) using the PyTorch version of the OpenNMT toolkit (Klein et al., 2017) and BPE tokenization (Sennrich, Haddow, and Birch, 2015) (joint vocabulary of 32K). The system was trained with 8.6M parallel sentences and evaluated on the FLORES-200 benchmark (Costa-jussà et al., 2022) obtaining 13.2 BLEU and 47.4 chrF++ for Spanish to Basque and 17.7 BLEU and 44.1 chrF++ for Basque to Spanish according to sacreBLEU tool (Post, 2018).

Paraphrasing language models: This approach is based on using pre-trained seq2seq language models trained on sentence paraphrases from synthetic paraphrase datasets. These paraphrase generators take a sentence as input and produce a set of paraphrased sentences. We have used two paraphrase generators, one for Spanish and one for Basque. The model used for Spanish is available in the Huggingface library and is trained on the Google PAWS dataset[3]. The paraphrase model for Basque has been generated by fine-tuning a BART model (Lewis et al., 2019) on a paraphrase dataset of 928 examples which is generated from combining the rewriting and synonym exercises of the EGA[4] exam we collected, and the Tapaco dataset[5], an automatically generated paraphrasing dataset from the tatoeba[6] parallel corpus. Since it does not exist a publicly available multilingual BART model including Basque, we pre-train a monolingual BART base[7] model (Lewis et al., 2019), trained on the same corpus that was used to train ElhBERTeu (Urbizu et al., 2022), and we apply BPE tokenization (Sennrich, Haddow, and Birch, 2015) (a vocabulary of 50K). We pretrain the model for 900K steps, with a batch-size of 256, and a sequence length of 128 tokens.

All the evaluated systems are based on fine-tuning chosen BERT models with the initial small manual seed dataset and the augmented set of examples achieved by means of the different paraphrasing strategies.

Table 6 and Table 7 show the results for the intent classification task on the augmented SportServ and FMTOD datasets, respectively, for the monolingual models. The augmentation of data has been performed in different proportions, where n indicates the number of synthetic examples added for each manual example on the dataset.

Lang.	Augment.	Back-trans.	P. Models
eu	n=1	76.19	70.56
eu	n=3	77.49	72.29
eu	n=5	**79.22**	71.48
eu	n=10	78.35	75.76
es	n=1	69.82	65.94
es	n=3	70.73	63.97
es	n=5	70.00	67.12
es	n=10	**75.68**	67.88

Table 6: Micro F1 results for the monolingual systems with data augmentation strategies on SportServ dataset. All the systems were evaluated with the pre-trained monolingual language models Berteus (Basque) and BETO (Spanish), depending on the language being tested (Lang.).

With the data augmentation strategy based on the back translation method, the intent classifier obtains a significant improvement in performance for both languages in the SportServ dataset. The best result is obtained by including five hypothetical translation examples in the data augmentation process for Basque and by including ten hypothetical translation examples for Spanish. It can be observed that the synthetic examples help to improve the baseline classifiers, both monolingual, which is improved by 3.5 points (F1=75.75 vs. F1=79.22) and multilingual, which is improved by almost 8 points (F1=71.43 vs. F1=79.22).

Table 7 shows that the synthetic examples help to improve the monolingual base-

[3] https://huggingface.co/mrm8488/bert2bert_shared-spanish-finetuned-paus-x-paraphrasing

[4] Certificate accrediting C1 of Basque

[5] https://huggingface.co/datasets/tapaco

[6] https://huggingface.co/datasets/tatoeba

[7] 12 layers: 6 encoder layers, 6 decoder layers.

Lang.	Augment.	Back trans.	P. Models
eu	n=1	90.04	87.88
eu	n=3	91.91	85.71
eu	n=5	91.34	87.88
eu	n=10	**92.12**	87.01
es	n=1	92.79	89.68
es	n=3	90.09	82.94
es	n=5	89.64	86.09
es	n=10	89.64	86.31

Table 7: Micro F1 results for the monolingual systems with data augmentation strategies on FMTOD dataset. All the systems were evaluated with the pre-trained monolingual language models Berteus (Basque) and BETO (Spanish), depending on the language being tested (Lang.).

line classifiers, but not the multilingual. For Basque the best result is obtained by including ten hypothetical translation examples in the data augmentation (F1=92.12) and outperforms the monolingual best baseline (F1=91.34). For Spanish, the best result is obtained by including one hypothetical translation example in the training (F1=92.79) and again, outperforms the best monolingual baseline (F1=90.54). As mentioned, the multilingual baselines are not surpassed by the monolingual data augmentation strategies on the FMTOD dataset.

In the experimentation, we have verified that the data augmentation strategy based on the paraphrase generation models is unable to outperform the baseline. In this case, the best result is achieved by adding ten paraphrases for each manual example in the case of SportServ dataset and by adding one paraphrase for each manual example in the case of FMTOD dataset, but, as mentioned, it does not improve the baseline.

7 *Multilingual intent-classification augmented with paraphrasis*

In this section we combine bilingual fine-tuning of the multilingual pre-trained language model mBERTeus with the back translation-based augmentation strategy. Due to the poor results obtained by the paraphrase generator models, we have only evaluated the back translation-based strategy in the multilingual systems.

In the case of the SportServ dataset for Basque, fine-tuning the multilingual BERT model with both, Basque and Spanish augmented data by means of back translation achieves the best results (F1=82.25). It outperforms the multilingual baseline (F1=71.43) by 11 points. The best results are obtained by adding ten hypothetical translation examples, but even the addition of a single example leads to improvements. In addition to all baselines in Table 4, it also overcomes the strategy of fine-tuning the monolingual pre-trained language model for Basque (Berteus) with augmented Basque data (see Table 6), even if with the initial manual data the best result was obtained with the monolingual model.

Test Lang.	Augmentation	Back translation
eu	n=1	77.06
eu	n=3	77.72
eu	n=5	80.52
eu	n=10	**82.25**
es	n=1	73.43
es	n=3	72.54
es	n=5	74.32
es	n=10	73.86

Table 8: Micro F1 results for the multilingual systems with data augmentation strategies on SportServ dataset.

On this same dataset for Spanish, combining the augmented data in both languages (F1=74.32) does not outperform the multilingual baseline (F1=74.67). In this case, the monolingual pre-trained model for Spanish BETO obtains the best results (see Table 6) when using only the augmented data in the target language for fine-tuning (F1=75.68).

Test Lang.	Augmentation	Back translation
eu	n=1	88.74
eu	n=3	90.48
eu	n=5	**93.08**
eu	n=10	90.04
es	n=1	91.44
es	n=3	91.44
es	n=5	91.91
es	n=10	88.99

Table 9: Micro F1 results for the multilingual systems with data augmentation strategies on FMTOD dataset.

In the case of the Basque FMTOD dataset, as in SportServ, the multilingual pre-trained language model fine-tuned with both, Basque and Spanish augmented data by means of back translation achieves the best results (F1=93.08).

Regarding Spanish, as is the case with the SportServ dataset, the multilingual synthetic data included in the fine-tuning process is not able to improve the performance of the multilingual baseline classifier (micro F1 score 91.44 vs 93.72). Furthermore, multilingual augmented systems perform below the results obtained with monolingual augmented systems.

8 Conclusions

When only a few manual examples exist, combining them to fine-tune a multilingual model is the most successful strategy, according to our experiments. However, it should be noted that in the case of Basque and the SportServ dataset, the monolingual model obtained better results than the multilingual training.

Monolingual data augmentation by paraphrasing using the back translation strategy is beneficial for both languages and in both datasets. In the case of the SportServ dataset results improve by 3.47 and 7.01 points for Basque and Spanish, respectively, and in the case of FMTOD, where the baseline is higher, results improve by around 1 point for Basque and 2 points for Spanish. Even in the case of Spanish, the strategy of fine-tuning the pre-trained monolingual models (Berteus and BETO) with the monolingual augmentation by means of back translation of the initial dataset outperforms the performance of the multilingual classifier fine-tuned with the data obtained with the multilingual augmentation. Nevertheless, combining both languages augmented examples obtained through the back translation strategy for fine-tuning the mBerteus multilingual language model acquire the best results in the case of the Basque language.

On the other hand, it has been proven that the zero-shot strategy is not feasible with so few manually annotated examples.

Furthermore, the paraphrase-generating language models have not been able to generate examples that improve the performance of the classifiers in either any language or any dataset. A future line of work is to understand the limitations of these models and to identify ways to enhance their effectiveness. In parallel, it would be interesting to analyse the capacity that the new Large Language Models, aligned by means of instruction fine-tuning and reinforcement learning, can offer in the paraphrase generation task.

Acknowledgements

This work has been partially funded by the Basque Government (ICL4LANG project, grant no. KK-2023/00094).

References

Adhikari, A., A. Ram, R. Tang, and J. Lin. 2019. 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)*, pages 4046–4051.

Agerri, R., I. S. Vicente, J. A. Campos, A. Barrena, X. Saralegi, A. Soroa, and E. Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of the 12th International Conference on Language Resources and Evaluation.*

Anaby-Tavor, A., B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, and N. Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020.*

Costa-jussà, M. R., J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672.*

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. 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)*, pages 4171–4186.

Dodge, J., G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Federmann, C., O. Elachqar, and C. Quirk. 2019. Multilingual whispers: Generating paraphrases with translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 17–26.

Goyal, T. and G. Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. *arXiv preprint arXiv:2005.02013*.

Jaderberg, M., V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, et al. 2017. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*.

Jolly, S., T. Falke, C. Tirkaz, and D. Sorokin. 2020. Data-efficient paraphrase generation to bootstrap intent classification and slot labeling for new features in task-oriented dialog systems. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 10–20, Online, December. International Committee on Computational Linguistics.

Klein, G., Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Kumar, V., A. Choudhary, and E. Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

López de Lacalle, M., X. Saralegi, and I. n. San Vicente. 2020. 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*, pages 2796–2802, Marseille, France, May. European Language Resources Association.

Mallinson, J., R. Sennrich, and M. Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893.

Meyer, S., D. Elsweiler, B. Ludwig, M. Fernandez-Pichel, and D. E. Losada. 2022. Do we still need human assessors? prompt-based gpt-3 user simulation in conversational ai. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–6.

Mosbach, M., M. Andriushchenko, and D. Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Otegi, A., A. Agirre, J. A. Campos, A. Soroa, and E. Agirre. 2020. Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 436–442.

Post, M. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ranta, A. 2004. Grammatical framework. *Journal of Functional Programming*, 14(2):145–189.

Schuster, S., S. Gupta, R. Shah, and M. Lewis. 2018. 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)*, pages 3795–3805, June.

Sennrich, R., B. Haddow, and A. Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909.*

Sokolov, A. and D. Filimonov. 2020. Neural machine translation for paraphrase generation. *arXiv preprint arXiv:2006.14223.*

Urbizu, G., I. San Vicente, X. Saralegi, R. Agerri, and A. Soroa. 2022. Basqueglue: A natural language understanding benchmark for basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, Y., C. Xu, Q. Sun, H. Hu, C. Tao, X. Geng, and D. Jiang. 2022. Promda: Prompt-based data augmentation for low-resource nlu tasks. *arXiv preprint arXiv:2202.12499.*

Wei, J. and K. Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196.*

Zhang, T., F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi. 2020. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987.*