Comparing model performance: Could TARS models be useful for low resource languages?

María Viana Rozas
Universidad del País Vasco/
Euskal Herriko Unibertsitatea
mviana009@ikasle.ehu.eus

David Lopez Loureiro
Universidad del País Vasco/
Euskal Herriko Unibertsitatea
dlopez105@ikasle.ehu.eus

Abstract

In this paper, we conduct a comparative analysis of the TARS and Flair models to assess their potential for application in low-resource languages. Through a comprehensive evaluation of the performance of these models in diverse language settings, we investigate the viability of employing TARS as a solution for low-resource languages. Our findings shed light on the extent to which TARS can effectively address the challenges associated with low-resource languages.

1 Introduction

In recent times, the field of artificial intelligence has made significant strides towards inclusivity by catering to languages with limited resources. However, despite these advancements, obtaining adequate datasets that yield reliable results remains a challenge. Often, researchers need to resort to multilingual datasets that may include the language of interest or rely on transfer learning from other languages during training.

In this study, our objective is to evaluate the performance of TARS models (Task Aware Representation of Sentences) with zero-shot classification in comparison to Flair-Embeddings, specifically in scenarios where data scarcity is a concern. We conduct extensive evaluations of these models in different languages, including Spanish, German, Japanese, and Basque, in order to assess their efficacy across languages with varying degrees of resource availability. This investigation enables us to glean insights into the performance of the system in languages with ample resources, limited resources, and those that are linguistically distinct from the Indo-European languages.

2 Related work

With the growing importance of natural language processing (NLP) in a wide range of applications, there has been increased interest in developing high-quality language models for low resource languages. However, due to the lack of training data and resources for these languages, developing effective models is challenging.

To address this issue, several studies have explored the use of transfer learning and pre-trained models to improve model performance on low resource languages. In particular, recent work has focused on the TARS model, a multi-task NLP model that has been shown to outperform other models on a variety of benchmark datasets.

In a study by (Halder et al., 2020), the authors compared the performance of TARS models and other state-of-the-art models on low resource languages, including Bengali, Swahili, and Yoruba. They found out that TARS models consistently outperformed other models on several metrics, including accuracy and F-score.

In (Jan Vium Enghoff and Agic, 2018) study about Low-resource NER they used 17 diverse languages with heterogeneous datasets and 2 massive parallel corpora. The results were that while standalone multisource annotation projection for NER can work when resources are rich in both quality and quantity, it is infeasible at a larger scale due to parallel corpora constraints.

As far as the development of this project is concerned, we have used as a basis for learning the presentation about NER, and all the papers related to it, like (Rodrigo Agerri, 2020), about Text Representation Models, in which they build monolingual word embeddings using FastText, and pre-trained language models using Flair and BERT.

3 Methodology

We employed two distinct yet related systems, namely TARS and Flair models, in our study.

- TARS Models: 'a novel formulation of text

classification to address crucial shortcomings of traditional transfer learning approaches' (Halder et al., 2020).

- Flair Embeddings: natural language processing (NLP) framework designed to simplify the training and distribution of advanced sequence labeling, text classification, and language models. The core concept of FLAIR is to provide a unified interface for different types of word and document embeddings, abstracting away the complexity of embedding-specific engineering. [1]

The overarching methodology involved multiple steps. Initially, we examined the performance of TARS models with Zero-Shot Classification. Subsequently, we trained the models with language-specific datasets, and finally, we evaluated the models using the same datasets to facilitate a comparative analysis of the results.

3.1 Zero-shot classification with TARS

TARS, introduced by (Halder et al., 2020), is a simple and effective method for few-shot and even zero-shot learning in text classification. This approach allows for text classification without the need for (m)any training examples.

One of the advantages of using TARS is that it allows users to input their own sentences for classification. However, a limitation of the model is that due to the lack of a large dataset, the learning can be challenging. Our studies have shown that TARS performs reasonably well with Germanic and Indo-European languages, such as German and Spanish, but is not as effective with more distant languages such as Basque and Japanese. This can be attributed to the fact that TARS models have been trained primarily on English and may not yield optimal results for other languages.

In our study, we used the zero-shot Named Entity Recognition (NER) tagger (tars-ner), which is trained primarily on English. We created sample sentences, annotated them with appropriate labels, and generated the corresponding NER tags using the tars-ner model.

3.2 TARS models

We trained separate TARS models for each language using different datasets. Due to the limited availability of labeled data for low-resource languages, using TARS alone did not yield satisfactory results. (Halder et al., 2020).

So, we decided to combine TARS models with Flair embeddings to improve the performance. In order to do that, we made use of TARSClassifier for transfer learning and FlairEmbeddings and WordEmbeddings for creating an embedding stack for the SequenceTagger, which is a regular Flair NER model. This combination allowed us to leverage the strengths of both TARS and Flair models for NER tasks in low-resource languages.

[1] See (Alan Akbik and Vollgraf, 2019) for more information about this model

Language	Dataset	Train Size	Labels
Spanish	Conll_03	8323	4
German	Conll_03 Biodiversity literature	12668	6
Basque	Ner_Basque	2297	4
Japanese	IOB2 Corpus	3621	17

Table 1: Comparison of the datasets for the different languages.

3.3 Flair models

We used the Flair lab 4 as a base for training Flair models for different languages. The same datasets used in the previous section were used to ensure performance comparison. We also employed the same embeddings and SequenceTagger as discussed earlier for consistency and comparability in our experiments.

4 Data

For our different approaches, we utilized the same datasets to enable accurate result comparison. Each language was associated with a specific dataset, which may not necessarily belong to the same domain (as shown in Table 1).

We determined the size of the datasets based on the training subset in its original format, without considering any downsampled versions that were used to expedite results during experimentation.

4.1 For Spanish

For Spanish, we selected the 'CONLL_03' dataset. It consists of two columns separated by a single space, with each word on a separate line and an empty line after each sentence. The first column contains words, and the second column contains named entity tags in the format used for the chunking task, with "B" denoting the first word of a named entity phrase and "I" denoting subsequent words. The named entity types include per-

son names (PER), organizations (ORG), locations (LOC), and miscellaneous names (MISC).

The dataset is sourced from news wire articles provided by the Spanish EFE News Agency and dates from May 2000. The dataset was annotated by the TALP Research Center of the Technical University of Catalonia (UPC) and the Center of Language and Computation (CLiC) of the University of Barcelona (UB).

The dataset contains a total of 8,323 sentences. However, for our work, we used a downsampled version to save time and GPU. resources.

4.2 For German

For German, we used the 'CoNLL-03 Biodiversity literature NER' dataset, which is based on multiple resources, including the BIOfid Corpus and an OCR parser. The BIOfid Corpus is a collection of historical scientific books focused on central European biodiversity, assembled by German domain experts. The OCR parser was used to extract text from these historical books.

The BIOfid corpus consists of approximately 15 journal titles with around 410 books, and 201 of these books (containing 969 articles) were selected by domain experts as a representative sample for generating training data for biological Named Entity Recognition (NER) [2]. The train set contains a total of 12,668 sentences. Similar to the previous dataset, we used a downsampled version, which should be taken into consideration when analyzing the results.

For labeling the named entities, the dataset uses the following categories: Time, Taxonomy, Location, Person, Organization, and Other, as it pertains to the domain of medicine.

4.3 For Basque

For Basque we used the 'NER_BASQUE' dataset, which is the first manually annotated corpus of Basque language specifically created for Named Entity Recognition (NER) task by the IXA Research Group. The dataset contains named entities classified into 4 categories: person (PER), location (LOC), organization (ORG), and other (OTH) named entities that do not belong to the previous 3 groups [3]. The train subset consists of a total of 2,297 sentences.

[2]For more information, see (S. Ahmed and Mehler, 2019)

[3]For more information, see (Iñali Alegria and Urizar, 2004)

Model	Language	Micro avg	Macro avg
TARS	Spanish	0.1434	0.6454
	German	0.4838	0.3690
	Basque	0.6207	0.4756
	Japanese	0.2933	0.1610
FLAIR	Spanish	0.8014	0.6916
	German	0.5968	0.4624
	Basque	0.6352	0.4869
	Japanese	0.3142	0.2015

Table 2: Comparison of the results obtained for the TARS models vs the Flair models based on micro and macro avg results of f-score.

4.4 For Japanese

For Japanese we used the 'IOB2 Corpus', which is based on the 'hironsan.txt' dataset containing Japanese language news articles from the Japanese version of Wikipedia. The dataset has been tokenized and annotated with IOB2 tags using MeCab, a Japanese morphological analysis tool. The corpus contains a total of 500 tagged sentences, and the tagging criteria are based on the definition of IREX (Information Retrieval and Extraction Exercise) with some approximations. [4]

The dataset has 17 labels, which is a relatively large number of labels, and this may have an impact on the performance of the trained models. The training subset consists of 3,621 sentences, and we decided not to use a downsampled version in order to fully make use of the available data.

5 Results

To evaluate the zero-shot classification, we used a qualitative approach based on human judgment. The results showed that for Spanish and German sentences, the system retains some labels but makes several mistakes, interpreting some book titles as authors. For languages further from English, such as Basque and Japanese, the system does not perform well and is unable to interpret any labels.

For quantitative comparison, we compared the performance of the TARS models with the Flair models only. Table 3 shows an example of the obtained results for Basque. We further present the results in terms of F-score for micro and macro averages in Table 2.

Checking the results based on macro and micro average can help interpreting the performance in different ways. The F-score micro can be useful for evaluating the overall performance of the NER

[4]For more information, see the github repository (Hironsan)

Model	Label	Precision	Recall	F-score
TARS	location	0.5987	0.5968	0.5978
	person	0.7425	0.6792	0.7094
	organization	0.6231	0.5700	0.5954
	other	0.0000	0.0000	0.0000
FLAIR	location	0.5468	0.7048	0.6158
	person	0.7519	0.6724	0.7099
	organization	0.6888	0.5666	0.6217
	other	0.0000	0.0000	0.0000

Table 3: Comparison of the results obtained for the TARS model of Basque vs the Flair model.

system in terms of its ability to identify all named entities in the dataset and the F-score macro can be useful for evaluating the performance of the NER system in terms of its ability to correctly identify all entity classes, regardless of their frequency in the dataset.

In addition to the results obtained, we compare the models in terms of their performance with example sentences.

6 Analysis

A more conclusive evaluation of the results is included in the notebook. Overall, looking at the different precision, recall, and F-score results for various languages, we can see that there are varying differences between the models.

However, the possibility of using zero-shot classification for languages far from English is not feasible at the moment, despite the authors' (Halder et al., 2020) intention to expand their resources in the future.

Regarding the TARS models, incorporating embeddings has significantly improved the performance compared to the results obtained without them, which were 0.0. By combining both possibilities, the results are closer to the Flair models, although they never surpass the performance of the latter.

Furthermore, the absence of embeddings for Basque and Japanese significantly reduces the performance of the models, as embeddings capture the context and semantics of words based on their context in a sentence. Since many named entities depend on the context in which they appear for their correct identification, the lack of this advantage results in poorer performance.

Regarding the performance of the models based on sentences, we can see that the Flair models outperform the TARS, although with minimal differences in languages such as Basque and Japanese.

7 Conclusions

Based on the analysis, the following conclusions can be drawn:

Zero-shot classification is not effective for languages other than English, especially for low-resource languages. The system struggles to accurately interpret named entities in non-Indo-European languages like Basque and Japanese.

Combining TARS with Flair embeddings improves the performance of the models, yielding results that are closer to those of Flair. However, using only TARS without embeddings is not viable, as the F-score average was very low (below 0.1). Therefore, the combination of TARS and Flair embeddings seems to be suitable, especially when dealing with datasets from low-resource languages. However, it is necessary to keep in mind that, although mixing the models improves the performance of TARS, actually, when combining the labels of the corpus with the newly defined ones, the system has to find the labels within that mix, which is a disadvantage.

Based on the comparison, if TARS is to be used for training, it is recommended to combine it with Flair embeddings to achieve better performance. Additionally, TARS shows relatively good results for small datasets, making it a potential method for low-resource languages when there is a lack of sufficient data.

For future research, it is recommended to develop embeddings for Japanese and Basque in order to improve the performance of the models in these languages.

In summary, while TARS models show decent performance for languages with more resources, it becomes more challenging for languages with fewer ones. However, by combining TARS with Flair embeddings and addressing the lack of embeddings for specific languages, it could be a viable option for low-resource languages in the future.

References

2005. Language-independent named entity recognition (i).

Duncan Blythe Kashif Rasul Stefan Schweter Alan Akbik, Tanja Bergmann and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp.

flairNLP. 2022. Few-shot and zero-shot classification (tars).

flairNLP. 2023. Loading a prepared corpus.

Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020. Task-aware representation of sentences for generic text classification.

Hironsan. Iob2 corpus.

Irene Balza Nerea Ezeiza Izaskun Fernandez Iñali Alegria, Olatz Agerri and Ruben Urizar. 2004. Design and development of a named entity recognizer for an agglutinative language.

Søren Harrison Jan Vium Enghoff and Zeljko Agic. 2018. Low-resource named entity recognition via multi-source projection: Not quite there yet?

M. F. Mbouopda and P. Melatagia Yonta. 2020. Named entity recognition in low resource languages using cross-lingual distributional word representation.

E. Munawwar. 2022. Zero and few shot learning - towards data science.

Jon Ander Campos Ander Barrena Xabier Saralegi Aitor Soroa Eneko Agirre Rodrigo Agerri, Inaki San Vicente. 2020. Give your text representation models some love: the case for basque.

Mitesh M. Khapra Rudra Murthy and Pushpak Bhattacharyya. 2018. Improving ner tagging performance in low-resource languages via multilingual learning.

C. Driller A. Pachzelt S. Ahmed, M. Stoeckel and A. Mehler. 2019. Biofid dataset: Publishing a german gold standard for named entity recognition in historical biodiversity literature.