UC3M at PAR-MEX@IberLef 2022: From Cosine Distance to Transformer Models for Paraphrase Identification in Mexican Spanish

Angela Brando-Le-Bihan[1,†], Roman Karbushev[1,†] and Isabel Segura-Bedmar[1]

[1] *Universidad Carlos III de Madrid, Leganés, 28914, Spain*

Abstract

In this paper, we describe the participation of the team UC3M for the task of Paraphrase Identification in Mexican Spanish (PAR-MEX@IberLEF 2022), which aims to identify if a sentence is a paraphrase of another sentence. We use Microsoft Research Paraphrase Corpus (MRPC) to tackle the task. In addition, we leverage several strategies such as class balancing or data augmentation to improve the generalization capability. Experimental results demonstrate the effectiveness of the MRPC approach (Microsoft Research Paraphrase Corpus), with an F1 score of 0.845 on the paraphrase class reported for the competition in the evaluation phase.

Keywords

NLP, Paraphrase Identification, Mexican Spanish, PAR-MEX 2022

1. Introduction

Paraphrase identification is an important task in Natural Language Processing (NLP) for supporting NLP applications such as plagiarism detection, machine translation or duplicate question matching on social media. This task aims to detect whether two sentence convey similar meanings [1].

This paper describes our participation in the competition "PAR-MEX@IberLef 2022: Paraphrase Identification in Mexican Spanish" [2] in the context of our Masters Degree in Science and Technology at UC3M.

The paraphrase task can be modeled as a binary classification aimed to identify whether two sentences have the same meaning or not. Thus, deep learning approaches usually achieve very successful results in this task [3, 4, 5].

However, most existing research has exclusively focused on the English language, while very few efforts have been dedicated to perform the task on other languages [6], mainly due to the lack of labeled and available datasets.

IberLEF 2022, September 2022, A Coruña, Spain.

[†] These authors contributed equally.

100476566@alumnos.uc3m.es (A. Brando-Le-Bihan); 100465163@alumnos.uc3m.es (R. Karbushev); isegura@inf.uc3m.es (I. Segura-Bedmar)
https://github.com/isegura/ (I. Segura-Bedmar)
0000-0002-7810-2360 (I. Segura-Bedmar)

The PAR-MEX@IberLef 2022 shared task aims to fill this gap by providing NLP researchers with an evaluation framework and a manually labeled dataset of sentence pairs written in Mexican Spanish. We analyzed several approaches to deal with this task in order to research and participate in this competition.

We summarize our contributions as follows:

- We analyze the training dataset and balance the classes with the techniques of over sampling and data augmentation.
- We explore different Machine Learning approaches for NLP tasks.
- The best approach for paraphrase identification was MRPC (Microsoft Research Paraphrase Corpus).

2. Methodology

2.1. Dataset Description

The training dataset provided for the competition contains 7,382 sentence pairs written Mexican Spanish, of which, only 17.37% (1,282 pairs) are paraphrases, that is, both sentences have similar meaning. Further details of the dataset are described at the competition's site PAR-MEX 2022. As we can see in Figure 1, the training dataset is highly unbalanced towards the negative class (that is, the sentence pairs that are not paraphrases). The unbalanced nature of the data is very common real world problems [7].

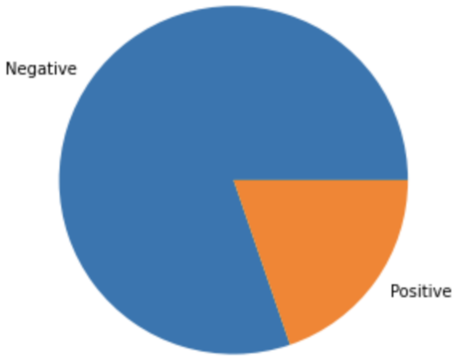

Figure 1: Class distribution in training dataset, where "Positive" refers to the distribution of the sentence pairs that are paraphrases.

2.2. Data sampling methods

To deal with the unbalanced dataset problem, we implement two different approaches. The first approach is based on oversampling, that is, duplicating the positive instances (sentence pairs labeled as paraphrases) to even out both classes. In particular, we use SMOTE [8] and

RandomOverSampler [9]. After applying these, the dataset is perfectly balanced. We also use a data augmentation approach based on generating paraphrases. To do this, we translate the sentence pairs classified as paraphrases to another languages, such as English, German or Chinese and, translate them back to Spanish. This translation cycle generates sentences that are slightly different from the original. To translate the sentences, we use the Google Translator libraries from deep_translator [10].

2.3. Classification Models

2.3.1. Cosine distance

The simplest approach is to calculate the cosine distance [11] between each possible sentence pair. The cosine distance can be regarded as a similarity metric between vectorized texts where, values close to 1 mean that both texts have very similar meanings (that is, they are paraphrases), while values close to 0 mean the opposite. Thus, as a previous step, we need to represent the sentences as vectors.

To vectorize the sentences, we use two very popular methods for text representation: the Tf-Idf model [12] (implemented with the scikit library) and the word embedding models provided by spaCy [13]. In particular, we use the es_core_news_lg model [14] for Spanish.

We calculate the cosine distance for all sentence pairs in the training dataset. This allows us to establish a threshold to properly classify each pair as paraphrase or not. After some testing, the optimal threshold was found to be 0.93. That is, a sentence pair will be classified as paraphrase only if its cosine distance is equal or greater than this threshold.

2.3.2. Traditional machine learning classifiers

We also exploit some classical machine learning classifiers such as Random Forest [15] and Support Vector Machine [16].

For text vectorization, we use the previously mentioned methods: spaCy vectors and tf-idf model. Also, we apply RandomOverSampler and SMOTE to balance the dataset. Several combinations were tested to find the most optimal. These will be described in Section 3.

2.3.3. MRPC (Microsoft Research Paraphrase Corpus)

The MRPC corpus [17] consists of 5,800 pairs of sentences from news in English. A team of human experts annotated each sentence pair as paraphrase (around 67%) or not (33%). It is available for download at The Microsoft website (last published: March 3, 2005).

The Hugging Face platform provides a pool of pre-trained models to perform a large variety of NLP tasks [18]. It provides APIs to download and experiment with the pre-trained models, with the possibility to fine-tune them with complementary datasets. In particular, we use the “bert-base-cased-finetuned-mrpc" model, which was fine-tuned for the task of paraphrase detection by using the MRPC corpus.

Given a sentence pair, the model provides a score indicating if both sentences convey the same meaning or not. For example, it can be used to check if the two sentences below are paraphrases:

- Sentence A = “sin embargo, para cuidar de la salud, se deberían comer ciertos alimentos y restringir otros cuantos."
- Sentence B = “el cuidado de la salud se debería de mantener comiendo ciertos alimentos y mantener restricciones con unos cuantos."

The model returns the probability for each class, which is the following for the sentences from the example:

- not paraphrase: 10%
- is paraphrase: 90%

That is, the two sentences are detected as a paraphrase with a probability of 90%.

3. Discussion and Results

Our baseline approach is based on cosine distance. As it was explained previously, we obtain a threshold that allows to identify the paraphrases. In this approach, we do not apply any data sampling or data augmentation techniques. Therefore, we provide the results for the unbalanced dataset. This approach achieves an accuracy of 0.85 over the test dataset. The cosine distance approach, although fast, was neither reliable nor good, performance-wise. Even though its accuracy was 0.85 on the test dataset, its performance was very poor when detecting positive (paraphrase) samples with only a 0.54 f1-score. Since it does not really require any training per se, a bigger dataset would not improve its performance.

In regards to traditional machine learning classifiers, SVM obtained the best performance. With the most optimal combination being: Tf-Idf for tokenization and RandomOverSampler for class balancing. This accomplished good model performance, evaluated on the test split from the train dataset provided by the organizers, with a 0.91 accuracy, 0.95 and 0.76 f1-scores for negatives and positives, respectively. This performance could be improved by a larger dataset since, as we can see in the learning curve plot from Figure 2, the model has yet to reach its full potential. Unfortunately, Spanish paraphrases corpora are very scarce.

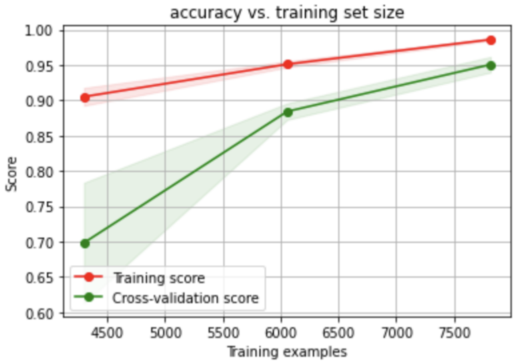

Figure 2: SVM + Tf-Idf + RandomOverSampler learning curve plot

Table 1 presents the results for the training, validation and test datasets for each traditional ML combination.

Table 1
Results provided by the Traditional ML models. F1-score is for positive class (paraphrases)

Model	Tokenization	Oversampling	N° sentences	F1 Score
Cos. Dist	spaCy	No	7,382	0.59
SVC	spaCy	No	1,477	0.29
SVC	spaCy	ROS	1,477	0.66
SVC	spaCy	SMOTE	1,477	0.63
SVC	Tf-Idf	No	1,477	0.42
SVC	Tf-Idf	ROS	1,477	0.79
SVC	Tf-Idf	SMOTE	1,477	0.64

Concerning the MRPC approach, we have adapted the code for evaluating each sentence using training, evaluation and test datasets, in order to classify each sentence pair as paraphrase or not. Table 2 presents the results for the training, validation and test datasets.

Table 2
Results provided by the MRPC model. F1-score is for positive class (paraphrases)

Phase	Number of sentences	F1 Score	Processing Time
Training	7,382	0.87	57 min. 31 sec.
Validation	97	0.8889	45 sec.
Test	2,819	0.845	22 min. 30 sec.

We must emphasize that, although the MRPC model was trained on news written in English, the model is quite effective for detecting paraphrases in sentences written in Mexican Spanish, with an appropriate F1-Score of 0.97 and 0.87 for negative and positive classes, respectively.

As we did not have more powerful machines, we used Google Collab for development. Unfortunately, we were not able to fine-tune this model with the training dataset provided by the organizers of the PAR-MEX shared task and in consequence we do not use data augmentation either for the MRPC approach. Processing was too heavy, because of the lack of resources (ex. GPU) and time was limited in the competition.

4. Conclusions

Paraphrase identification is a complex problem that requires a lot of data pre-processing and model fine tuning to achieve acceptable performance.

In our research, we found that oversampling methods greatly improve model performance of traditional machine learning classifiers such as SVM. Even though these classifiers can demonstrate good performance, large datasets and great computational power are needed to achieve that. Unfortunately, the creation of these datasets is very expensive and time-consuming. Thus, the lack of annotated datasets is one of the main bottlenecks for supervised machine learning algorithms applied to NLP tasks. Our best traditional machine learning approach, Tf-Idf + SVM, showed good performance with the training dataset (0.91 accuracy) but a poor one (0.342 score) on the final test dataset for the competition.

Fortunately, we can take advantage of transfer learning, by using pre-trained models and later fine-tuning to a specific task such as paraphrases detection. The Hugging face platform already provides many of these models. In particular, we have used a BERT model fine-tuned with the MRPC corpus for the paraphrase detection task. This model obtains an F1 of 0.845 on the test dataset, which is an improvement of 5 points over the results provided by SVM. Our team ranked 6th in the competition.

As future work, we plan to research on data augmentation techniques and explore multilingual approaches for dealing with the tasks in different scenarios. We will try to create datasets for the task in different Spanish languages. We will also fine-tune our own model for paraphrase detection.

5. Acknowledgments

This work was supported by the Community of Madrid under the Interdisciplinary Projects Program for Young Researchers (NLP4Rare-CM-UC3M project) and the line of Excellence of University Professors (EPUC3M17). It was also funded by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain, (ACCESS2MEET project-PID2020-116527RB-I00).

References

[1] R. Mihalcea, C. Corley, C. Strapparava, et al., Corpus-based and knowledge-based measures of text semantic similarity, in: Aaai, volume 6, 2006, pp. 775–780.

[2] G. Bel-Enguix, H. Gomez-Adorno, G. Sierra, J.-M. Torres-Moreno, J.-G. Ortiz-Barajas, J. Vasquez, Overview of PAR-MEX at Iberlef 2022: Paraphrase Detection in Spanish Shared Task, Procesamiento del Lenguaje Natural 69 (2022).

[3] H. Wang, F. Ma, Y. Wang, J. Gao, Knowledge-guided paraphrase identification, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 843–853.

[4] A. Nighojkar, J. Licato, Improving paraphrase detection with the adversarial paraphrasing task, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 7106–7116. URL: https://aclanthology.org/2021.acl-long.552. doi:10.18653/v1/2021.acl-long.552.

[5] Z. Shi, M. Huang, Robustness to modification with shared words in paraphrase identification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 164–171. URL: https://aclanthology.org/2020.findings-emnlp.16. doi:10.18653/v1/2020.findings-emnlp.16.

[6] Y. Yang, Y. Zhang, C. Tar, J. Baldridge, PAWS-X: A cross-lingual adversarial dataset for paraphrase identification, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong

Kong, China, 2019, pp. 3687–3692. URL: https://aclanthology.org/D19-1382. doi:10.18653/v1/D19-1382.

[7] A. Mountassir, H. Benbrahim, I. Berrada, An empirical study to address the problem of unbalanced data sets in sentiment classification, in: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2012, pp. 3298–3303. doi:10.1109/ICSMC.2012.6378300.

[8] I. learn, Over-Sampling methods - Smote, 2022. https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html.

[9] I. learn, Over-Sampling methods - RandomOverSampler, 2022. https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html.

[10] N. Baccouri, https://deep-translator.readthedocs.io/en/latest/usage.html, 2020.

[11] A. Kumar, Cosine Similarity and Cosine Distance, 2020. https://medium.datadriveninvestor.com/cosine-similarity-cosine-distance-6571387f9bf8.

[12] scikit learn, Feature Extraction - Tf-Idf Transformer, 2022. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html.

[13] spaCy, SPACY, 2022. https://spacy.io.

[14] spaCy, SPACY - SPANISH, 2022. https://spacy.io/models/es.

[15] T. Yiu, Understanding Random Forest, 2019. https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

[16] A. Yadav, SUPPORT VECTOR MACHINES, 2018. https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589.

[17] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, CoRR abs/1804.07461 (2018). URL: http://arxiv.org/abs/1804.07461. arXiv:1804.07461.

[18] bert-base-cased-finetuned-mrpc, https://huggingface.co/bert-base-cased-finetuned-mrpc, 2020. Accessed: 2022-05-28.