Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 255-263 recibido 05-07-2022 revisado 18-07-2022 aceptado 21-07-2022

Overview of PAR-MEX at Iberlef 2022: Paraphrase Detection in Spanish Shared Task

Resumen de PAR-MEX en IberLEF 2022: Tarea Compartida para la Detección de Paráfrasis en Español

Gemma Bel-Enguix[1], Gerardo Sierra[1], Helena Gómez-Adorno[2], Juan-Manuel Torres-Moreno[3], Jesus-German Ortiz-Barajas[4], Juan Vásquez[4]
[1] Instituto de Ingeniería (UNAM)
[2] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (UNAM)
[3] Laboratoire Informatique d'Avignon (Avignon Université)
[4] Posgrado en Ciencia e Ingeniería de la Computación
{gbele,gsierram}@iingen.unam.mx, helena.gomez@iimas.unam.mx,
juan-manuel.torres@univ-avignon.fr, {jgermanob,juanmv}@comunidad.unam.mx

Abstract: Paraphrase detection is an important unresolved task in natural language processing; especially in the Spanish language. In order to address this issue, and contribute to the creation of high-performance paraphrase detection automated systems, we propose a shared task called PAR-MEX. For this task, we created a corpus, in Spanish, with topics in the domain of Mexican gastronomy. Afterwards, the participants in this task submitted their classification results on our corpus. In this paper we explain the steps followed for the creation of the corpus, we summarize the results obtained by the various participants, and propose some conclusions regarding the paraphrase-detection task in Spanish.
Keywords: PAR-MEX, paraphrase detection, Iberlef.

Resumen: La detección de paráfrasis es una tarea importante no resuelta en procesamiento del lenguaje natural; especialmente en la lengua española. Para atacar este problema, y para contribuir a la creación de sistemas de detección automática que obtengan resultados competitivos, proponemos la tarea compartida llamada PAR-MEX. Para esto, creamos un corpus en español con temas dentro del campo semántico de gastronomía mexicana. Después los participantes en esta tarea enviaron los resultados de sus sistemas de clasificación sobre nuestro corpus. En este paper explicamos los pasos seguidos para la creación del corpus, resumimos los resultados obtenidos por los participantes, y proponemos algunas conclusiones al respecto de la detección de paráfrasis en español.
Palabras clave: PAR-MEX, detección paráfrasis, Iberlef.

1 Introduction

Two texts, or two sentences, are paraphrase when they are semantically equivalent, regardless of the cause that led to that equivalence (Das and Smith, 2009). Detecting paraphrased text is a task that has aroused the interest of the Natural Language Processing (NLP) community, due to the fact that it has multiple applications, such as plagiarism detection, question-answering and machine translation (Kong et al., 2020).

Paraphrase construction includes different mechanisms, such as lexical changes through synonymy, sentence rearrangement, breaking of a sentence into several parts, and joining more than one phrase into another. Therefore, addressing the problem of paraphrase detection requires an analysis that encompasses different levels, both lexical and semantic, as well as syntactic.

To deal with the problem of paraphrase detection using supervised machine learning methods, researchers use data sets that typically include pairs of sentences that are identified as paraphrase or non-paraphrase. There are various ways of elaborating or compiling these corpora: news collections, plagiarism pairs, manual creation, relational ac-

ISSN 1135-5948. DOI 10.26342/2022-69-22

Topic of the document	No. of lines
sushi	28
molecular cuisine	21
tequila	25
kebab	25
day of the dead	25
vegan food	25
street food	25

Table 1: Topic and number of lines in each of the seven original documents.

quisition, back-translation, multiple translations.

The PARMEX task has been organized for the first time at IberLeF 2022 (Montes-y-Gómez et al., 2002), a shared evaluation campaign for NLP systems in Spanish and other iberian languages, which is part of the SEPLN congress. The task is based on the Gastronomy Corpus, elaborated by the Language Engineering Group, which is divided into seven sub-corpora that deal with different topics related to cuisine, preferably, but not exclusively, Mexican. The corpus has been manually compiled in Mexico and, therefore, contains some terms and expressions specific to the Mexican variant of Spanish.

The rest of their paper is organised as follows. Section 2 presents the evaluation framework used at PARMEX 2022. Section 3 shows an overview of different approaches taken to tackle the problem. Section 4 reports and analyses the results obtained by the teams that have participated. Finally, Section 5 presents our conclusions from this shared task.

2 PARMEX 2022 Corpus and evaluation framework

For the PAR-MEX at Iberlef 2022 task, we created a corpus comprised of sentence pairs in Mexican Spanish. For the creation of the sentence pairs, first we produced seven original texts with gastronomical topics. Each one of these seven texts had a variable number of lines. On Table 1 the exact number of lines and topics per document are shown.

The second step in the creation of the corpus was the generation of the paraphrased documents. These new documents were created by humans who were tasked with writing one document with identical semantic content and same number of lines as in the original document. For example, for the document `sushi.txt`, an original document with 28 lines, seven paraphrased documents were created. The 28 lines in each one of these seven paraphrased documents contained the exact same meaning as the 28 lines in the original document.

Topic	No. of paraphrased documents
sushi	7
molecular cuisine	31
tequila	7
kebab	7
day of the dead	8
vegan food	6
street food	6
$\sum$	**72**

Table 2: Topics of the original seven documents, and their respective number of paraphrased documents.

The process described above was repeated for every one of the seven original documents. Then, we generated a total of 72 paraphrased documents. The exact numbers can be seen on Table 2.

The next step in the elaboration of the task's corpus was the creation of the sentence pairs, and their respective labels. For this, we paired each line in every original document with each line in every paraphrased document. If the sentence pair was made up of a line in an original document with an index of i, and one line in a paraphrased document with an index i, then it would be labeled as "paraphrase". In the opposite case, the one in which a sentence in the original document with index i was matched with a sentence from another document but with an index of j (given that $i \neq j$), then that sentence pair would be labeled as "not paraphrase". It is important to mention that even if the index of an original document and the index of a paraphrased document were equal, it was also verified that the line from the paraphrased document belonged to the same topic as the line from the original document. For example, if line i from document `vegan_food.txt` was paired with line i from a paraphrased document related to `tequila.txt`, this pair would not be labeled as paraphrase since their semantic contents would differ due to their topics even though their indices were

Topic	No. of high-level sentence-pairs
sushi	41
molecular cuisine	214
tequila	84
kebab	63
day of the dead	75
vegan food	42
street food	51
$\sum$	**750**

Table 3: Number of high-level paraphrase pairs per original document.

the same. Therefore, in order to obtain the paraphrase sentence-pairs, the topic and the indices were compared.

The final step in the creation of the corpus was the addition of the high-level paraphrase pairs. For this, we requested humans to write several original documents with high-level paraphrase. During this step, we did not ask them to write paraphrased documents with the same number of lines as the original documents. Once created these novel documents with high-level paraphrases, we extracted some lines and paired them with the sentences in the original documents. This process generated less paraphrase pairs than the initial step with low-level paraphrases, and the exact number of high-level paraphrase-pairs can be observed in Table 3.

After the pairing of the sentences, and the creation of their respective labels, a total of 10,298 sentence-pairs were obtained. From this set, 1,844 sentence-pairs were labeled as paraphrase, while the remaining 8,454 sentence-pairs were labeled as non-paraphrase. This represented an approximate of 20% of sentence-pairs labeled as paraphrase, with the remaining 80% labeled as not paraphrase. From this set, we created the training, validation and test partitions. The distribution of these sets is shown on Table 4.

3 *Overview of the Submitted Approaches*

In this edition, six teams submitted one or more solutions to the task through the codalab platform[1]. CodaLab Competitions is a robust open-source framework for running competitions that involve results or code submission. The evaluation methodology of a competition in this platform consists of receiving as input the predictive outputs of systems. It returns a performance evaluation based on the metrics defined for each task.

Partition	Total sentence-pairs	Paraphrase sentence-pairs
Training	7,382	1,282
Validation	97	20
Test	2,819	542
Total	10,298	1,844

Table 4: Number and distribution of sentence-pairs in the training, validation and evaluation sets.

This section presents a summary of the submitted systems in terms of preprocessing, feature extraction, and classification algorithms. In Table 5 we indicate the general approach used by each team. It can be appreciated that participants used two general approaches: transformers and traditional ML. Following this, we briefly describe each of the participants methods.

Approach	NLP-CIC-TAGE	Tü-Par	Thang CIC	Abu	FRSCIC	UC3M-DEEPNLP
Transformers	X		X			X
Traditional ML		X		X	X	

Table 5: General approach of each participating team.

- *Using Transformers on Noisy vs. Clean Data for Paraphrase Identification in Mexican Spanish* (Tamayo, Burgos, and Gelbukh, 2022)
  - **Team name: NLP-CIC-TAGE**
  - **Summary:** The participants presented a transfer learning approach using transformers to tackle paraphrase identification on noisy vs. clean data in Spanish. They used BERTIN, a pre-trained model on the Spanish portion of a massive

[1] https://codalab.lisn.upsaclay.fr/competitions/2345

multilingual web corpus. The fine-tuning and parameter tunning of BERTIN was performed on noisy data and used to identify paraphrase on clean data.

- *PAR-MEX Shared Task Submission Description: Identifying Spanish Paraphrases Using Pretrained Models and Translations* (Girrbach, 2022)
  - **Team name: Tü-Par**
  - **Summary:** The participants proposed an approach based on a classical machine learning pipeline consisting of feature extraction, supervised learning, and evaluation. The feature extraction consists in encoding Spanish sentences (or their English translations) by a pretrained sentence encoder, then concatenating the sentence embeddings or representing the sentences by a similarity score. Different classifiers were used depending on the feature type—a logistic regression model and a random forest model on the similarity features, and multi-layer perceptrons on the sentence embeddings features.

- *GAN-BERT, an Adversarial Learning Architecture for Paraphrase Identification* (Ta et al., 2022)
  - **Team name: Thang CIC**
  - **Summary:** The participants used text embeddings from pre-trained transformer models for training by GAN-BERT, adversarial learning. They modified noises for the generator, which have a random rate and the exact size of the hidden layer of transformers. They also included a rule of thumb based on the pair similarity to remove possible wrong sentence pairs in positive examples and additional unlabelled data in the same domain to improve the model performance.

- *Paraphrase Identification: Lightweight effective methods based features from pre-trained models* (Rahman et al., 2022)
  - **Team name: Abu**
  - **Summary:** The participants introduced two lightweight methods: linear regression and multilayer perceptron, trained on six features: the difference in sentences' length, common lemmas between 2 sentences, sentences' similarity, etc. After performing Component Analysis (PCA) to reduce the dimension, they filter noises in the positive examples by introducing a rule of thumb on the pair similarity.

- *Mexican Spanish Paraphrase Identification using Data Augmentation* (Meque et al., 2022)
  - **Team name: FRSCIC**
  - **Summary:** The participants performed a data augmentation step on the training set using translation. The text vectorization process consisted of sentence transformers, spaCy vectors, traditional word n-grams, and bi-tri syntactic n-grams using TF-IDF. They proposed a similarity vector using three different similarity algorithms for the final representation: Jaccard, Cosine, and spaCy. For the classification step, they used a soft-voting ensemble model with three estimators.

- *UC3M at PAR-MEX@IberLef 2022: From Cosine Distance to Transformer Models for Paraphrase Identification in Mexican Spanish* (Brando-Le-Bihan, Karbushev, and Segura-Bedmar, 2022)
  - **Team name: UC3M-DEEPNLP**
  - **Summary:** The participants evaluated a baseline method based on the cosine similarity of two text pairs representation: TF-IDF model on bag-of-words and word embedding models provided by spaCy. For the final submission, they used the "bert-base-cased-finetuned-mrpc" model, which is

fine-tuned for paraphrase detection by using the MRPC corpus. They also proposed strategies such as class balancing or data augmentation to improve the generalization capability. However, they did not present these strategies in the final submission.

4 Experimental Evaluation and Analysis of the Results

This section reviews the results obtained by the participants of PAR-MEX at Iberlef 2022: Paraphrase Detection in Spanish Shared Task. For this purpose, we analyse and compare the submitted solutions' performance on the test partition. We used the F1-score metric on the paraphrase (P) as the primary performance measure and to rank all the participants. We launched a Codalab competition to manage the shared task stages and compute the performance metric for all submissions.

We propose a transformers-based approach as a baseline. It consists of the Bidirectional Encoder Representation from Transformer (BERT) model (Devlin et al., 2019). We use the base model for our baseline, consisting of twelve Transformer blocks and the pre-trained model BETO (Cañete et al., 2020), a BERT model trained on an enormous Spanish corpus. We use four epochs and the Adam optimizer for the fine-tuning stage with a learning rate of 2e-5. We use the HuggingFace implementation (Wolf et al., 2020) for Tensorflow (Abadi et al., 2015). In order to have comparable results with the participant submissions, we report the best result in five runs using different random seeds.

Table 6 summarises the results obtained by each team and our baseline in the PAR-MEX shared task. We report the F1 score in both Paraphrase and Non-paraphrase classes, the macro F1 score, and the accuracy. In this edition of the PAR-MEX shared task, the approach submitted by the NLP-CIC-TAGE team outperformed all the other approaches and the baseline. The NLP-CIC-TAGE team used an approach based on a transformer architecture; they fine-tuned the RoBERTa model pre-trained in a Spanish corpus. In contrast, the second-best approach proposed by the Tü-Par team used a Random Forest classifier using similarity-based features. These results show that classic approaches are still competitive for this task compared to deep learning.

We use the Maximum Possible Accuracy (MPA) and Coincident Failure Diversity (CFD) metrics (Tang, Suganthan, and Yao, 2006) to analyse the complementariness and the diversity of the predictions of the submitted approaches. The MPA is analogous to accuracy, defined as the correct classified instances divided into the total number of instances. To consider an instance correctly classified, at least one of the teams needs to assign the correct label to it. Using the MPA metric, we can detect the misclassified instances by all teams. The CFD metric has a minimum value of 0 when all classifiers are always correct or when all classifiers are either correct or wrong. On the other hand, it has a maximum value of 1 when at most one classifier will fail on any randomly chosen instance (Kuncheva and Whitaker, 2003). The CFD is defined in equation 1.

$$CFD = \begin{cases} 0, & p_0 = 1.0; \\ \frac{1}{1-p_0} \sum_{i=1}^{L} \frac{L-i}{L-1} p_i & p_0 < 1 \end{cases} \tag{1}$$

Table 7 shows the results of these metrics by grouping the proposed approaches based on their similar features. We create four groups: all teams, all teams who send their paper, Transformers-based approaches, traditional-machine-learning-based approaches. All of the groups mentioned above have at least two members. All participants sent their papers but one. Transformers-based approaches include the following teams: NLP-CIC-TAGE, Thang CIC, and UC3M-DEEPNLP. Tü-Par, FRSCIC, and ThangCIC conform traditional machine learning based approaches group.

In terms of the general approach, traditional machine learning performs better in terms of MPA than Transformers-based solutions. The above suggests that the different features used to train these machine learning models complement each other. In the same way, the combination of transformers and machine learning approaches obtain the highest MPA performance and have an average increment of 0.66% compared to those

Team	F1-score (P)	F1-score (NP)	Accuracy	Macro F1-score
NLP-CIC-TAGE	0.9424	0.9869	0.9787	0.9647
Tü-Par	0.9373	0.9853	0.9762	0.9613
Thang CIC	0.9022	0.9775	0.9635	0.9399
Abu	0.8867	0.9751	0.9592	0.9309
FRSCIC	0.8754	0.9730	0.9557	0.9242
UC3M-DEEPNLP	0.8450	0.9679	0.9468	0.9065
temu_bsc	0.8441	0.9567	0.9322	0.9004
baseline	0.834936	0.953075	0.926924	0.894006

Table 6: Result summary for the PAR-MEX shared task on the test set.

Approach	Best accuracy	MPA	CFD	Number of systems
All teams	0.9787	0.9936	0.0408	7
all teams (with submission)	0.9787	0.9915	0.0341	6
Transformers	0.9787	0.9847	0.0331	3
Traditional ML	0.9762	0.9883	0.0373	3

Table 7: MPA and CFD comparison results among the different proposed approaches.

individual approaches. Finally, the values for the CFD score are comparable among all approaches, which means that their predictions are complementary to an extent; this leads us to conclude that traditional and transformer-based approaches learn different information from text pairs.

Table 8 shows the results of the F1 score for the paraphrase class divided by topic in the test set. The kebab category achieved the highest performance with an average F1 score of 0.9483; on the other hand, the sushi topic had the worst performance with an average F1 score of 0.7659. The NLP-CIC-TAGE team obtained the best performance in two of the seven topics. In contrast, the Tü-Par team obtained the best performance in four topics, including the sushi, which is the hardest. Nevertheless, the difference was in the food truck topic. The NLP-CIC-TAGE obtained a 0.9153 F1 score, while the Tü-Par team obtained 0.8673. For this result, the NLP-CIC-TAGE achieved first place in the PAR-MEX shared task.

Tables 9 and 10 show the performance of each team by topic and low-level paraphrase and high-level paraphrase, respectively. In order to compute these metrics, we filtered the paraphrase examples and kept the non-paraphrase examples unchanged. Only day of the dead, vegan food, and food truck topics have examples of high-level paraphrase. Regarding high-level paraphrase, the food truck topic obtains the highest performance while the sushi topic obtains the lowest; however, the sushi topic only has one example of this type of paraphrase. When comparing high-level and low-level paraphrase performance, only the food truck topic performs better on high-level paraphrase than on low-level paraphrase. These results suggest that, in general, detecting high-level paraphrase examples is more challenging for the proposed approaches. The most substantial difference is in the vegan food topic; the average result in high-level paraphrases is 0.6509, while in low-level paraphrases is 0.9095, which means a 0.2586 between both levels. This topic has 41 high-level paraphrase examples and 36 low-level paraphrase examples; because the examples of this topic are nearly balanced, we can conclude that the performance difference is due to the difficulty of identifying high-level paraphrase features.

In terms of proposed approaches, Transformers-based models outperform all teams in two of the three topics with high-level paraphrase examples; in the remaining topic, Transformers-based and traditional machine learning approaches have the same performance. Therefore, we can conclude that Transformers can learn better features to identify high-level paraphrases. On the other hand, when dealing with low-level paraphrases, a traditional machine learning approach outperform all teams in 4 of 7 topics. A Transformers-based approach has the highest performance in the remaining

Team	Molecular cusine	Day of the dead	Kebab	Tequila	Vegan food	Sushi	Food truck
NLP-CIC-TAGE	0.9878	0.9714	0.9792	0.9231	0.8261	0.8333	0.9153
Tü-Par	0.9762	0.9859	0.98	0.9362	0.8252	0.8772	0.8673
Thang CIC	0.9687	0.8400	0.9216	0.9091	0.8444	0.7692	0.8468
Abu	0.9495	0.9489	0.9574	0.8864	0.8000	0.6567	0.7573
FRSCIC	0.9254	0.8806	0.9293	0.8764	0.7852	0.8077	0.7810
UC3M-DEEPNLP	0.9010	0.8000	0.9462	0.8989	0.7576	0.6818	0.7358
temu_bsc	0.8460	0.8675	0.9245	0.7132	0.8591	0.7353	0.9167
Baseline	0.7871	0.9863	0.8596	0.7833	0.8387	0.7692	0.918
Average	0.9177	0.9096	0.9372	0.8658	0.8181	0.7654	0.8412

Table 8: Results for the PAR-MEX shared task on the test set by topic.

Team	Molecular cusine	Day of the dead	Kebab	Tequila	Vegan food	Sushi	Food truck
NLP-CIC-TAGE	0.9878	0.9636	0.9792	0.9231	0.9333	0.8511	0.9189
Tü-Par	0.9762	1	0.98	0.9362	0.9114	0.8727	0.8406
Thang CIC	0.9687	0.8099	0.9216	0.9091	0.9577	0.7843	0.806
Abu	0.9495	0.9541	0.9574	0.8864	0.9429	0.6667	0.6885
FRSCIC	0.9254	0.8571	0.9293	0.8764	0.8919	0.8	0.7302
UC3M-DEEPNLP	0.901	0.7473	0.9462	0.8989	0.8919	0.6977	0.6769
temu_bsc	0.846	0.8382	0.9245	0.7132	0.9	0.7273	0.9067
Baseline	0.7871	0.9828	0.8596	0.7833	0.8471	0.7619	0.9091
Average	0.9177	0.8941	0.9372	0.8658	0.9095	0.7702	0.8096

Table 9: Results for the PAR-MEX shared task on the test set by topic and low-level paraphrases.

three topics. With these results, we can conclude that machine learning models can handle low-level paraphrasing better than complex models like transformers when using similarity-based features as the primary type of characteristics.

Finally, Table 11 shows each team's performance only on the paraphrase type. Again, the results are consistent with what we show in tables 7 and 8. Although the NLP-CIC-TAGE team does not obtain the best result in every topic in the test set, their overall performance is the best on both levels of paraphrasing.

5 Conclusions

This paper described the design and results of the PAR-MEX shared task collocated with IberLef 2022. PAR-Mex is focused in paraphrase identification in Mexican Spanish texts. This has been the first edition of the task.

The data set of PAR-MEX included both, low-level and high-level pairs of paraphrases, although they were not distinguished for the participants. The analysis of the results shows that, whereas low-level paraphrase is currently an easy task for natural language processing (0.90 of average), high-level paraphrase is a problem that has not been conveniently approached yet.

The best results in this shared task were obtained by a team that proposed to approach the problem with a method based on transformers. However, traditional machine learning strategies obtained very similar results. Indeed, while deep learning techniques have the best scores in the sub-corpora of molecular cuisine, vegan food sushi and food truck, traditional methods lead in day of the dead, kebab and tequila. The only topic in which transformers reach a clearly better score is food truck. This shows this is a complex task and that collaboration between models and the use of multiple variables can improve the final outcome of the research.

Acknowledgments

We acknowledge the support of the projects CONACyT CB A1-S-27780, and DGAPA-UNAM PAPIIT references TA400121 and TA101722. The authors thank CONACYT for the computing resources provided through the Deep Learning Platform for Lan-

Team	Day of the dead	Vegan food	Sushi	Food truck
NLP-CIC-TAGE	1	0.6567	0	0.9091
Tü-Par	0.9286	0.6479	0.2222	0.9091
Thang CIC	0.6364	0.7077	0	0.9091
Abu	0.9286	0.623	0	0.8571
FRSCIC	0.875	0.6061	0.25	0.8571
UC3M-DEEPNLP	0.9655	0.5397	0	0.7727
temu_bsc	0.5769	0.7273	0.1	0.913
Baseline	0.9375	0.6988	0.1176	0.8936
Average	0.8561	0.6509	0.0862	0.8776

Table 10: Results for the PAR-MEX shared task on the test set by topic and high-level paraphrases. Molecular cuisine, kebab and tequila do not have high-level paraphrase examples.

Team	F1-score high-level paraphrase	F1-score low-level paraphrase
NLP-CIC-TAGE	0.7755	0.9602
Tü-Par	0.6951	0.9538
Thang CIC	0.6552	0.9137
Abu	0.6494	0.905
FRSCIC	0.6795	0.8884
UC3M-DEEPNLP	0.6575	0.8605
temu_bsc	0.4167	0.8378
Baseline	0.3976	0.8265
Average	0.6158	0.8927

Table 11: Results for the PAR-MEX shared task on the test set by paraphrase type.

guage Technologies of the INAOE Supercomputing Laboratory.

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Brando-Le-Bihan, A., R. Karbushev, and I. Segura-Bedmar. 2022. UC3M at PAR-MEX@IberLef 2022: from cosine distance to transformer models for paraphrase identification in mexican spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*.

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*.

Das, D. and N. A. Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 468–476, Suntec, Singapore, August. Association for Computational Linguistics.

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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Girrbach, L. 2022. PAR-MEX shared task submission description: Identifying spanish paraphrases using pretrained models and translations. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022).*

Kong, L., Z. Han, Y. Han, and H. Qi. 2020. A deep paraphrase identification model interacting semantics with syntax. *Complexity*, 2020:14 pages.

Kuncheva, L. and C. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207, 05.

Meque, A., F. Balouchzahi, G. Sidorov, and A. Gelbukh. 2022. Mexican spanish paraphrase identification using data augmentation. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022).*

Montes-y-Gómez, M., J. Gonzalo, F. Rangel, M. Casavantes, M. Álvarez-Carmona, G. Bel-Enguix, H. Escalante, L. Freitas, A. Miranda-Escalada, F. Rodríguez-Sánchez, A. Rosá, M. Sobrevilla-Cabezudo, M. Taulé, and R. Valencia-García. 2002. *Proceedings of IberLeF 2002.*

Rahman, A., H. Ta, L. Najjar, and A. Gelbukh. 2022. Paraphrase identification: Lightweight effective methods based features from pre-trained models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022).*

Ta, H., A. Rahman, L. Najjar, and A. Gelbukh. 2022. GAN-BERT, an adversarial learning architecture for paraphrase identification. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022).*

Tamayo, A., D. A. Burgos, and A. Gelbukh. 2022. Using transformers on noisy vs. clean data for paraphrase identification in mexican spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022).*

Tang, E. K., P. N. Suganthan, and X. Yao. 2006. An analysis of diversity measures. *Machine learning*, 65(1):247–271.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.