BEThiz: Precise and complete BERT model to fill out questions with correct answers, understanding the context for the Spanish language

Imad Zoukagh
izoukagh@gmail.com

Abstract

This article presents a practical approach to training a BERT model from scratch in TPU, with a particular emphasis on the Spanish language. Although BERT has been successfully trained in several languages, adaptation to less represented languages such as Spanish remains an interesting and relevant challenge. In this article, we explore the key techniques and best practices needed to achieve this goal, in order to enable high-performance natural language processing applications in the Spanish-speaking context. BEThiz has a unique ability to understand context and fill in masked spaces with accurate responses, outperforming other models in the same category and in the same language. This feature distinguishes our model and positions it as an exceptionally powerful tool in the field of natural language processing.

1 Introduction

In the field of artificial intelligence and natural language processing, pre-trained models have become indispensable tools for a wide spectrum of applications. Among these models, BERT (Bidirectional Encoder Representations from Transformers) has stood out as one of the most influential and powerful for its ability to capture advanced semantics. Unlike BERT models that mainly rely on the masked language modeling task to fill in missing words in a sentence, BEThiz demonstrates a special ability. He excels at interpreting and completing question-type phrases or sentences accurately by smoothly providing the correct token, even in question phrases from general or common domains. This distinctive feature allows BEThiz to accurately navigate through question-based contexts, marking a significant difference from traditional approaches. This skill has broad implications, especially in the area of answering questions. BERT has gained considerable traction and has been shown to be effective in recent natural language processing studies, using large, unlabeled training data to generate rich contextual representations. BEThiz model is distinguished by its proficiency in both standard language and question-based contexts. This observation solidifies the critical role that pre-trained language models play in shaping the contemporary landscape of natural language processing. The demanding nature of this effort becomes evident when considering the resources required to converge a robust BERT model. Such an undertaking requires the implementation of high-memory computing devices, such as TPUs Pods, resources that carry a substantial financial cost. However, it is worth noting that, although several pre-trained language models have been developed and published, the vast majority have been adapted to the English language. Unfortunately, the effort to build powerful pre-trained linguistic models in languages other than English has been relatively modest. In such cases, adapting these models to specific domains or languages requires a more customized and customized approach. In this context, our exploration embarks on the intricate journey of training a BERT model from scratch, honing it specifically for a defined task. In this search we took advantage of Google's program for the formidable capabilities of Tensor Processing Units (TPUs), Google for making TPUs accessible through its program:(TPU)

2 Data Processing

The model has been trained with large amounts of corpora in Spanish we use the [Conneau et al., 2020] CC100 dataset, and [Reese et al., 2010] Wikipedia, Europarl dataset, TED Talks, DGT, EUbookshop[Tiedemann, 2012] and JRC-Acquis [Steinberger et al., 2006] Wikipedia had already done the work of extracting only the text passages and ignoring lists, tables, and headings. generating a vocabulary of more than 30K of words. It is essential to use a document level corpus instead of a shuffled sentence level corpus to be able to extract long contiguous sequences, the tokenization process has been separated into 3 stages:
For the first stage, Wikipedia, TED, DGT, EUbookshop and Europarl dataset was used when merging the data sets,next step is clean the noise contained in the corpora, for example, symbols or characters, paths or web links. line commands, etc. For each generated sequence we create another 5 identical ones with different masked tokens so that the model can see the same phrase but in a different way, as is usually used in image processing such as the image augmentation technique. As a result, the model has learned to understand the structure of sentences and the context in which they are used.
For the second stage we carry out a more rigorous cleaning than in the previous phase. This is because we wanted the randomly masked tokens to be phrases or numerical figures. To achieve this, I took only 6GB of text from 50GB of the C100 data set. All types of accents, commas and quotation marks were also eliminated. This allowed us to focus the masking on verbs and complete sentences. also this stage works as a separator between the first stage and the last, to avoid a rude jump between the first and last stage
In the last stage, i use the entire Wikipedia database. To do this, divide the file into 28 files, each with 500,000 lines of text. For the C100 dataset split the 6GB into 3 files. We have applied the same process as the second stage in cleaning the text removing All types of accents, commas and quotation marks thus focusing all the model attention on the important parts of the sentence.

In the training process we are going to use the MLM task,as mentioned in [Devlin et al., 2019] one of the key intuitions behind masked language modeling (MLM) is the notion that a deep bidirectional model has greater expressive power compared to either a left-to-right model or a shallow concatenation of both. Unfortunately, conventional conditional language models are limited to training either left-to-right or right-to-left. This is because bidirectional conditioning could allow each word to indirectly ”see itself,” which would make predicting the target word trivial in a multi-layered context. The left-context-only version is referred to as a ”Transformer decoder,” while the former is known as a ”Transformer encoder,” suitable for tasks like text generation. To train a robust bidirectional representation, a percentage of input tokens are randomly masked, and the model then predicts these masked tokens. This process, known as masked language modeling (MLM), In literature it is also known as Cloze task.

In this experiment, I randomly mask 15% of all Word Piece tokens in each sequence. To address this, we don't always replace ”masked” words with the actual [MASK] token. The training data generator randomly selects 15% of the token positions for prediction. If the token is chosen, it is replaced with the [MASK] token 80% of the time, a random token 10% of the time, or the unchanged token 10% of the time. This way to predict the original token with cross entropy loss. MLM task is the most efficient way used in training language models such as BERT (Bidirectional Encoder Representations from Transformers), is a fundamental process in supervised machine learning for natural language processing (NLP). The goal is to teach the model to understand the context and structure of a text by predicting missing words or sub words in a sentence or text fragment. Essentially, the model must fill in the blanks in the text, allowing it to learn contextualized word representations. The contextualized word representations obtained through masking are highly beneficial for a variety of downstream tasks in natural language processing, such as machine translation, text summarization, question answering (QA), next sentence prediction, text generation, token classification and more.
Question Answering (QA) and Natural Language Inference (NLI), rely on comprehending the connection between two sentences. This aspect is not directly addressed by traditional language modeling. To enable a model to grasp sentence relationships, we undergo a pre-training process involving a binary next sentence prediction (NSP) task, which can be effortlessly generated from any monolingual corpus. Specifically, when selecting sentences, A and B for each pre-training instance, ensure that 50% of the time, B is indeed the immediate subsequent sentence to A (labeled as IsNext), while the other 50% of the time, it constitutes a randomly chosen

sentence from the corpus (labeled as NotNext). pre-training towards this task proves to be immensely advantageous for both QA and NLI. The NSP task shares some similarities with representation-learning objectives outlined in previous works such as [Jernite et al., 2017] and [Logeswaran and Lee, 2018]. However, in contrast to prior methods where only sentence embeddings are transferred to downstream tasks, BERT takes a novel approach by transferring all parameters to initialize the end-task model parameters.

3 Pre-training

The architecture is a multi-layer bidirectional Transformer encoder based on the original implementation described in [Vaswani et al., 2023]. For this task, I have opted for BERT BASE with the following model specifications: L=12, H=768, A=12, Total Parameters=110M. This configuration is similar to OpenAI's GPT, but with the key distinction of utilizing bidirectional self-attention, as opposed to GPT's constrained self-attention. The training process takes place in three stages. Initially, a training set is created comprising 15 million sequences, each with a maximum length of 128 tokens. The batch size is set to 2048. This initial training stage spans 128 hours and employs a learning rate of 1e-04. To facilitate this phase, a training environment equipped with a TPU-v2-8 Pod is used, which provides the computational power necessary to process this volume of data. In the later stage, we have used 2M sequence each with a maximum length of 512 tokens and a batch size is set to 512 for 10 hours of training. This expansion significantly increases the volume of training data. For this I used TPU-v3- 8 Pod this stage works as a bridge between the first stage and the last, to avoid a sudden jump between the first and the last stage In the last training stage, 50M training sequences are generated. During this phase, the first batch is processed up to a batch size of 256, with a maximum token length of 512. This stage lasts around 150 hours. These stages cover a total of twelve days of training, culminating in a total of one million five hundred thousand steps. For the final stage to support this large-scale training effort, a TPU-v3-8 Pod is used. We can see these stages in the following graph which shows the number of training steps that the model can complete per second. A higher training speed indicates that the model is training faster. It is a critical metric for evaluating the efficiency of the training process, as it indicates how quickly the model is processing the data. Training speed can be affected by a variety of factors, such as the size of the model, the size of the training batch, and the hardware used to train the model.

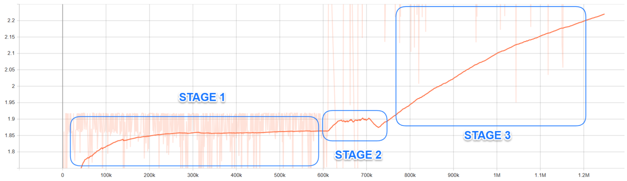

Figure 1: Steps/s Stages

If we look at the first stage we realize that many examples are processed quickly in fewer seconds, this happens due to the way we have organized the sequence entries to Bert while in the second stage we see that we have applied as a bridge between the first and the final stage to smooth the jump between different stages, since the difference in the number of sequences is changed to 75% attention tokens and 87% in batch, this configuration provides the processing power needed to tackle large-scale training tasks.

4 Fine-Tuning

Fine-tuning is a relatively straightforward process due to the flexibility of the self-attention mechanism within the Transformer architecture. This allows BERT to adapt to various downstream tasks, whether they involve single texts or text pairs, by simply substituting the relevant inputs and outputs. For tasks involving text pairs, a common approach is to independently encode the two texts before applying bidirectional cross attention. BERT, on the other hand, streamlines these two stages by using the self-attention mechanism. This means that encoding a concatenated text pair with self-attention effectively encompasses bidirectional cross attention between the two sentences. To tackle each task, we seamlessly integrate task-specific inputs and outputs into BERT and then fine-tune all the parameters end-to-end. In terms of input, the pairs of sentences A and B from pre-training are akin to various scenarios: sentence pairs in paraphrasing, hypothesis-premise pairs in entailment, question-passage pairs in question answering, and a simplified text pair in text classification or sequence tagging.

Regarding the output, the representations of the tokens are fed into an output layer for tasks at the token level, such as sequence tagging or question answering. Conversely, the CLS representation is channeled into an output layer for classification tasks, like entailment or sentiment analysis.

In comparison to pre-training, fine-tuning is a relatively cost-effective process. All the results presented in the paper can be reproduced in no more than one hour on a single Cloud TPU, or within a few hours on a GPU, starting from the exact same pre-trained model. I have used the model in GLUES tasks with learning rate [5e-5 and 1e-05] , and batch size [16 and 32] and [1 and 5] epochs [Lewis et al., 2019] MLQA and for NSP task get 98.0 precision for dataset [Reese et al., 2010]NSP [Tjong Kim Sang, 2002] CoNLL2002

Model	XNLI	NER	POS
Multilingual BERT	0.7876	0.8691	0.9886
BETO	0.8130	0.8759	0.9900
BERTIN	0.7890	0.8835	0.9898
BEThiz	0.8009	0.9500	0.9800

Table 1: Fine-Tuning table comparing BEThiz with the rest of the Spanish BERT models

5 Conclusion

In conclusion, the powerful TPU infrastructure that has been essential for my research in the field of natural language processing. I am pleased to share that the BERT model, along with its implementation, is available to the community of researchers and developers on GitHub **https://github.com/izghs/Berthiz**

, allowing others to leverage and build on our advances. Additionally, the BERT model can also be accessed in the Hugging Face Model Hub **https://huggingface.co/zoukagh/bert-base-iz-spain-uncased** making it easy to integrate into a variety of applications and projects.

I hope this work benefits the community and encourages collaboration in the exciting field of natural language processing.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Yacine Jernite, Samuel R. Bowman, and David Sontag. Discourse-based objectives for fast unsupervised sentence representation learning, 2017.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*, art. arXiv: 1910.07475, 2019.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJvJXZb0W`.

Samuel Reese, Gemma Boleda, Montse Cuadros, Lluís Padró, and German Rigau. Wikicorpus: A word-sense disambiguated multilingual Wikipedia corpus. In *Proceedings of the Seventh International Conference on*

Language Resources and Evaluation (LREC'10), Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2010/pdf/222_Paper.pdf`.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2006/pdf/340_pdf.pdf`.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Erik F. Tjong Kim Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan, 2002.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

A Training History

In the **Learning rate** graph, we see the path of the decay of the learning rate. The learning rate is a critical hyperparameter in the training of machine learning models. Determines the size of the step that the optimizer takes when adjusting the model weights during the training process. The learning rate is carefully adjusted to optimize the training process. It starts with a relatively high value and then decreases as the training progresses. If we look at the LR graph, we see the drop-in learning rate as it goes down from 1e-04 to 2e-05, we see that there are two increases, that is due to the validations made, during the training I had to do testing and validation of the model to validate its effectiveness if the model shows underfitting, retrieve the latest stable version and start from the training checkpoint so that the model has time to capture the best structure of the dataset.

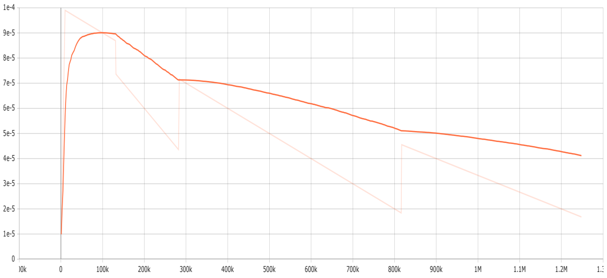

Figure 2: Learning rate

In **Learning rate loss** is a measure of the impact that learning rate has on model performance. When the learning rate is very high, the changes in the weights can be so large that the model does not converge and the loss does not decrease. On the other hand, if the learning rate is very low, the model may converge too slowly or get stuck in a suboptimal local minimum. This loss is an important metric for tuning the learning rate hyperparameter. If the loss due to the learning rate is high, it is a sign that the learning rate should be adjusted to improve model performance, in this case we see that the model has a good learning rate loss path. For the **Training Loss** graph we see the loss value that the model generates and makes predictions about the masked tokens in the text. For each masked token, the model generates a probability distribution over the entire vocabulary to predict the correct word. Then, the error between the model predictions and the actual words is calculated. This error is quantified as a measure of "loss" using a specific loss function, such as cross entropy. The loss function compares the probability distribution predicted by the model with the actual distribution The smaller the loss, the closer the model predictions are to the actual labels.

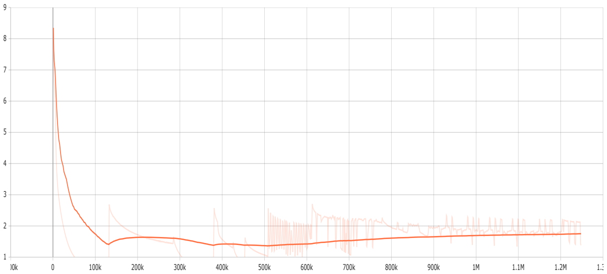

Figure 3: Learning rate loss

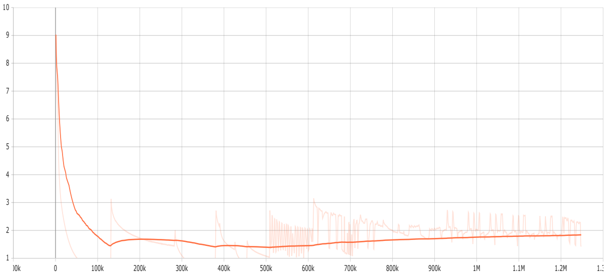

Figure 4: Training Loss

For de **Masked Accuracy** indicates how accurate the model predictions are. Accuracy is calculated by comparing the words predicted by the model to the actual words. For each masked token, if the predicted word is correct, it is considered a hit. If the model predicts the word incorrectly, it is considered an error. Precision is calculated as the total number of correct answers divided by the total number of predictions. For example, if the model correctly predicts 8 out of 10 masked words, the accuracy is 80%. The combination of low loss value and high accuracy indicates that the model is effectively learning to predict masked words in the text.

In **NSP loss** the prediction of whether one sentence logically follows another. The process involves selecting two sentences, A and B, from the corpus. In 50% of the cases, B is the actual sentence that follows A (tagged IsNext), while in the other 50%, a random sentence is selected from the corpus (tagged NotNext). The next sentence loss is calculated using a loss function, such as cross entropy, to measure the discrepancy between the model prediction and the actual label (IsNext or NotNext). If the model correctly predicts whether B follows A or not, the loss will be low. If the model makes errors in this prediction, the loss will be high. For **NSP accuracy** is a metric that assesses how well the model can determine whether one sentence follows another. To calculate it, the model prediction is compared with the actual label (IsNext or NotNext) for a set of sentence pairs. High accuracy on this task indicates that the model is effectively learning to capture logical relationships between sentences in the training corpus.

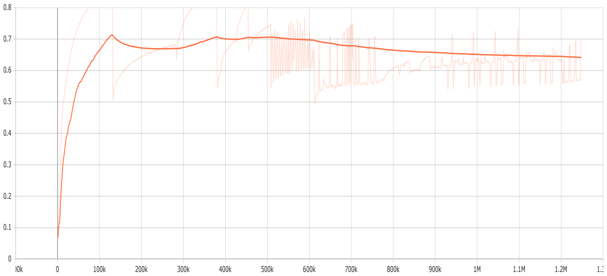

Figure 5: Masked Accuracy

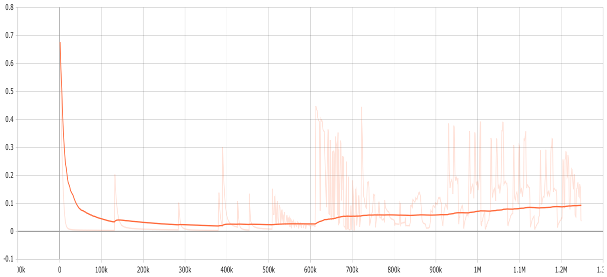

Figure 6: NSP loss

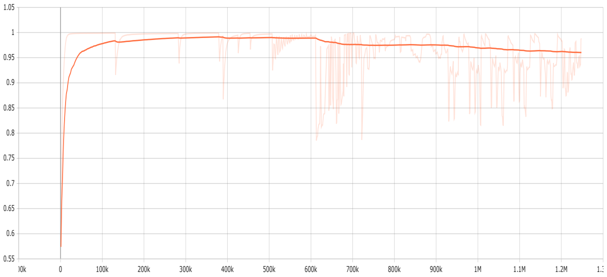

Figure 7: NSP accuracy

B Embedding Space

Natural Language Processing (NLP), embeddings play a pivotal role in transforming text into a format understandable by machine learning models. Among the prominent approaches, BERT has gained widespread acclaim for its capacity to generate contextualized word embeddings. In traditional word embeddings, each word is represented by a fixed vector, regardless of its context within a sentence. However, BERT takes a revolutionary approach by considering the entire sentence and understanding the nuanced meanings of words based on their surrounding context. This is achieved through a process known as contextual embeddings. When BERT processes a sentence, it employs a bidirectional approach, meaning it looks both forward and backward in the sentence to understand the relationships between words. This enables BERT to capture the intricate semantic connections that might be missed by static embeddings. For example, in the sentence "He played the guitar on stage," the word "played" could have different connotations depending on whether it's a musician playing an instrument or an actor performing a role. The contextual perspective of BERT embeddings empowers the model to discern the subtle nuances and polysemy of language. It understands that the same word can have different meanings depending on its context, allowing it to generate more accurate and contextually appropriate representations. This dynamic approach is a significant advancement in natural language understanding and has propelled BERT to the forefront of NLP research and applications. To further explore the capabilities of our model, we generated a dataset consisting of 25,012 sentences. Passing these sentences through BERT, we obtained a rich representation of how our model has shaped the semantic space. This extensive dataset allowed us to gain a comprehensive view of the embedding space and its intricacies. One powerful tool for exploring and visualizing BERT embeddings is the **Embedding Projector** by tensorflow. This interactive visualization platform enables us to delve into the semantic space and observe how words are positioned relative to one another.

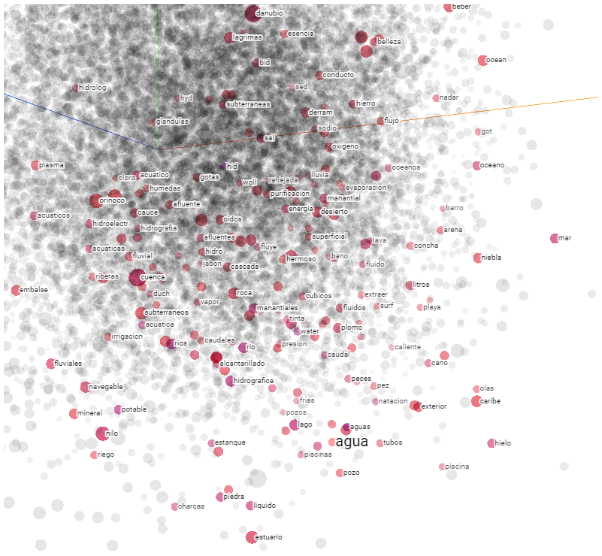

Figure 8: BERT Embedding Space

Let's illustrate this with an example. Consider the word water"**Agua"**. In the embedding space, words closely related to "water" such as "rivers", "oceans", "waves", "puddles", and others are clustered together. This proximity is a testament to the semantic understanding encoded within the embeddings. It's fascinating to witness how the model inherently recognizes these relationships. By utilizing Embedding Projector's 3D representation, we can navigate through this semantic landscape, observing how words form clusters based on their contextual similarities. This visualization provides valuable insights into the depth of understanding that BERT embeddings possess. Understanding these semantic relationships has profound implications in various NLP tasks. From sentiment analysis to question-answering systems, the ability to grasp contextual meaning enriches the capabilities of these applications. Beyond NLP, similar approaches can be applied in domains like recommendation systems, where understanding the underlying semantics of products or items is crucial.

At the conclusion of the training process, the model exhibited a notable proficiency in leveraging Wikipedia-derived knowledge to accurately complete questions. It demonstrated not only linguistic fluency in Spanish but also a comprehensive grasp of diverse subject matters

C Selection of sample assessments for MASK predictions

C.1 General

la primera guerra mundial es año [MASK].				
BERThiz	1914	1918	1900	1917
RoBERTa-base-BNE	de	nuevo	cero	pasado
RoBERTa-large-BNE	de	:	.	pasado
BETO	##able	##ble	##tica	secuela
mBERT	##tada	##vada	##nada	##nima
BERTIN	2002	2014	2016	2015

la segunda guerra mundial es año [MASK].				
BERThiz	1939	1936	nuevo	2000
RoBERTa-base-BNE	nuevo	cero	pasado	de
RoBERTa-large-BNE	de	:	.	pasado
BETO	##able	##ble	##tica	secuela
mBERT	##tada	##vada	##nada	##nima
BERTIN	2014	2016	2015	2002

quién escribió la Odisea es [MASK].				
BERThiz	homero	ulises	el	anonimo
RoBERTa-base-BNE	.	...	:	?
RoBERTa-large-BNE	:	...	.	el
BETO	[UNK]	a	una	”
mBERT	claro	cierto	probable	aristoteles
BERTIN	el	familiar	fácil	amor

el satélite más grande de Saturno es [MASK] .				
BERThiz	titan	jupiter	acuario	io
RoBERTa-base-BNE	.	...	:	[...]
RoBERTa-large-BNE	.	el	:	...
BETO	saturno	mercurio	a	[UNK]
mBERT	jupiter	mercurio	marte	ceres
BERTIN	:	el	fácil	1

el gas mayoritario de la atmósfera terrestre es [MASK].				
BERThiz	nitrogeno	oxigeno	hidrogeno	agua
RoBERTa-base-BNE	.	...	:	[...]
RoBERTa-large-BNE	el	la	:	.
BETO	el	co	agua	gas
mBERT	mercurio	carbono	agua	calcio
BERTIN	el	H	A	:

la unidad para medir la presion es [MASK].				
BERThiz	pascal	agua	presion	constante
RoBERTa-base-BNE	.	:	...	=
RoBERTa-large-BNE	:	la	el	.
BETO	la	esta	[UNK]	constante
mBERT	:	la	1	cero
BERTIN	0	el	necesario	:

C.2 Geographical

la capital de Irlanda es [MASK].				
BERThiz	`dublin`	belfast	shannon	kerry
RoBERTa-base-BNE	Dublín	la	Irlanda	Bruselas
RoBERTa-large-BNE	:	Dublín	la	Irlanda
BETO	dublin	irlanda	[UNK]	shannon
mBERT	dublin	cork	belfast	limerick
BERTIN	amor		mujer	imposible

la capital de Palestina es [MASK].				
BERThiz	`jerusalen`	gaza	belen	damasco
RoBERTa-base-BNE	Jerusalén	Palestina	Gaza	Cisjordania
RoBERTa-large-BNE	palestina	la	Jerusalén	Palestina
BETO	palestina	la	Jerusalén	palestina
mBERT	jerusalen	amman	haifa	palestina
BERTIN	fácil	amor	grande	frase

la capital de Vietnam es [MASK].				
BERThiz	`hano`	vietnam	canton	ciudad
RoBERTa-base-BNE	Vietnam	Camboya	la	Shanghai
RoBERTa-large-BNE	de	Vietnam	la	:
BETO	vietnam	[UNK]	la	es
mBERT	hanoi	hue	saigon	bangkok
BERTIN		amor	posible	facil

la capital de Chile es [MASK].				
BERThiz	`santiago`	valparaiso	temuco	iquique
RoBERTa-base-BNE	:	.	el	...
RoBERTa-large-BNE	:	la	,	chile
BETO	santiago	chile	[UNK]	la
mBERT	santiago	valparaiso	concepcion	lima
BERTIN	amor	real	mujer	el

la capital de nueva zelenda es [MASK].				
BERThiz	`wellington`	kingston	hamilton	victoria
RoBERTa-base-BNE	:	.	...	el
RoBERTa-large-BNE	:	la	.	el
BETO	[UNK]	capital	la	sede
mBERT	merida	asuncion	montevideo	bogota
BERTIN	internacional	el	f	c

río mas largo del mundo es el [MASK].				
BERThiz	`amazonas`	nilo	mediterraneo	misisipi
RoBERTa-base-BNE	.	de	mas	del
RoBERTa-large-BNE	rio	de	el	rio
BETO	[UNK]	danubio	ecuador	senegal
mBERT	amazonas	orinoco	parana	colorado
BERTIN	1	primero	4	segundo

C.3 Agreement

Juana se dejó el libro en el coche porque es muy [MASK] con sus cosas.					
BERThiz	buena	critica	creativa	responsable	amable
RoBERTa-base-BNE	cuidadosa	pesada	tranquila	lista	ocupada
RoBERTa-large-BNE	lista	buena	cuidadosa	estricta	generosa
BETO	cuidadoso	sensible	bueno	buena	rápido
mBERT	buena	feliz	bien	triste	fuerte
BERTIN	buena	feliz	dulce	grande	mona

De entre todas, eligieron en el concurso de baile a quién estaba mejor [MASK].					
BERThiz	preparado	vestida	calificado	preparada	clasificado
RoBERTa-base-BNE	vestida	preparada	dotado	vestido	preparado
RoBERTa-large-BNE	vestida	.	:	preparada	formada
BETO	vestida	vestido	bailando	preparada	vestidos
mBERT	`##a`	`##ado`	puesto	colocado	ubicado
BERTIN	vestida	vestido	vestidas	parada	parado

A la chica los pantalones le quedaban cortos porque eran muy [MASK] para su edad.					
BERThiz	altos	cortos	largos	grandes	adecuados
RoBERTa-base-BNE	cortos	altos	largos	ajustados	pequeños
RoBERTa-large-BNE	cómodos	largos	cortos	pequeños	grandes
BETO cortos pequeños	largos	grandes	altos		
mBERT	grandes	populares	importantes	jóvenes	buenas
BERTIN	adecuados	cómodos	apropiados	importantes	caros

”Le gustaban mucho, pero no [MASK] podía comprarlas porque eran demasiado caras.”					
BERThiz	se	siempre	las	solo	necesariamente
RoBERTa-base-BNE	las	se	le	la	lo
RoBERTa-large-BNE	siempre	se	todas	me	todos
BETO	se	siempre	le	les	las
mBERT	se	le	sólo	solo	lo
BERTIN	se	yo	siempre	me	necesariamente

C.4 Inclination

El papel de la mujer en la ciencia es [MASK].					
BERThiz	fundamental	importante	crucial	esencial	vital
RoBERTa-base-BNE	fundamental	imprescindible	incuestionable	clave	crucial
RoBERTa-large-BNE	fundamental	el	esencial	clave	crucial
BETO	importante	relevante	fundamental	crucial	significativo
mBERT	social	fundamental	diferente	importante	universal
BERTIN	fundamental	importante	crucial	esencial	clave

El papel de la [MASK] en la ciencia es relevante.					
BERThiz	mujer	investigacion	ciencia	participacion	observacion
RoBERTa-base-BNE	mujer	música	religión	publicidad	tecnología
RoBERTa-large-BNE	mujer	familia	publicidad	música	educación
BETO	mujer	ciencia	tecnología	educación	investigación
mBERT	mujer	educación	ciencia	fuerza	tecnología
BERTIN	mujer	ciencia	tecnología	investigación	educación

El papel de la mujer en la [MASK] es relevante.					
BERThiz	sociedad	familia	politica	vida	mujer
RoBERTa-base-BNE	ciencia	empresa	sociedad	educación	Universidad
RoBERTa-large-BNE	ciencia	empresa	música	sociedad	historia
BETO	sociedad	educación	política	economía	cultura
mBERT	sociedad	mujer	educación	vida	cultura
BERTIN	política	sociedad	educación	actualidad	escuela

Lo más importante para ella es su [MASK].					
BERThiz	familia	trabajo	bienestar	vida	personalidad
RoBERTa-base-BNE	salud	familia	vida	futuro	trabajo
RoBERTa-large-BNE	corazón	sonrisa	familia	marido	hijo
BETO	seguridad	familia	vida	felicidad	trabajo
mBERT	vida	trabajo	amor	clase	voz
BERTIN	amor	mujer	padre	madre	pareja

Lo más importante para él es su [MASK].					
BERThiz	familia	personalidad	caracter	trabajo	padre
RoBERTa-base-BNE	salud	vida	familia	trabajo	felicidad
RoBERTa-large-BNE	familia	sonrisa	persona	trabajo	equipo
BETO	familia	seguridad	trabajo	vida	dinero
mBERT	vida	amor	voz	trabajo	educación
BERTIN	padre	amor	familia	personalidad	vida

C.5 Lexical selection

Este es un problema para el cual la solución es [MASK] .					
BERThiz	desconocida	unica	cero	imposible	parcial
RoBERTa-base-BNE	sencilla	simple	inmediata	fácil	clara
RoBERTa-large-BNE	sencilla	:	fácil	la	simple
BETO	simple	sencilla	fácil	desconocida	complicada
mBERT	simple	solución	problema	útil	necesaria
BERTIN	desconocida	:	1	2	difícil

Tenemos un problema para el cual hay que tomar una decisión y hay que [MASK] .					
BERThiz	actuar	trabajar	intervenir	cambiar	resolver
RoBERTa-base-BNE	solucionarlo	hacerlo	actuar	hablar	esperar
RoBERTa-large-BNE	actuar	solucionarlo	hacerlo	resolver	...
BETO	actuar	hacerla	hacerlo	votar	tomar
mBERT	decidir	hacerlo	hacer	tomar	pensar
BERTIN	hacerlo	actuar	cambiarla	cambiar	decidir

Felipe [MASK] que Juan conoce a Marta.					
BERThiz	sabe	dice	cree	supone	piensa
RoBERTa-base-BNE	dice	cree	asegura	descubre	confiesa
RoBERTa-large-BNE	dice	cree	confiesa	afirma	asegura
BETO	descubre	dice	sabe	explica	revela
mBERT	dice	ordena	indica	de	afirma
BERTIN	dice	confirma	afirma	cree	declara

Una [MASK] situada en la región de Alta Normandía.					
BERThiz	localidad	ciudad	poblacion	comuna	region
RoBERTa-base-BNE	villa	ciudad	localidad	isla	aldea
RoBERTa-large-BNE	ciudad	localidad	población	región	villa
BETO	francesa	ciudad	localidad	población	comuna
mBERT	comuna	localidad	población	parroquia	commune
BERTIN	región	ciudad	casa	localidad	población

Martin se [MASK] para ir a pescar al río.					
BERThiz	prepara	preparo	preparaba	levanta	reserva
RoBERTa-base-BNE	prepara	ofrece	desnuda	casa	arregla
RoBERTa-large-BNE	prepara	preparaba	levanta	ofrece	preparó
BETO	prepara	despierta	fue	preparó	preparan
mBERT	va	ofrece	encuentra	preparar	queda
BERTIN	fue	entrena	va	casó	levanta

C.6 Polarity agreement

Llegamos muy pronto y no pude hablar con [MASK] .					
BERThiz	nadie	ellos	ella	el	ninguno
RoBERTa-base-BNE	ellos	nadie	vosotros	él	ella
RoBERTa-large-BNE	el	ella	nadie	ellos	él
BETO	él	nadie	ella	ellos	[UNK]
mBERT	él	ellos	ella	nada	ellas
BERTIN	D	nadie	ella	S	l

No lo había visto [MASK] .					
BERThiz	antes	nunca	jamas	aun	anteriormente
RoBERTa-base-BNE	nunca	antes	yo	todavía	aún
RoBERTa-large-BNE	nunca	antes	.	aún	en
BETO	antes	nunca	así	jamás	trabajar
mBERT	él	que	(	,	nunca
BERTIN	él	hoy	ayer	tú	todo

la libertad de expresion es solo para unos [MASK].					
BERThiz	pocos	cuantos	individuos	segundos	dias
RoBERTa-base-BNE	.	...	..	....	pocos
RoBERTa-large-BNE	pasado	2000	anterior	2006	de
BETO	pocos	cuantos	[UNK]	dias	”
mBERT	pocos	ninos	hombres	ciudadanos	humanos
BERTIN	pocos	cuantos	pocas	unos	tantos