Proceedings of the Workshop on Hybrid Intelligence for Natural Language Processing Tasks HI4NLP (co-located at ECAI-2020)
Santiago de Compostela, August 29, 2020, published at http://ceur-ws.org

Departamento de Nosotros: How Machine Translated Corpora Affects Language Models in MRC Tasks

Maria Khvalchik[1] and **Mikhail Galkin** [2]

Abstract. Pre-training large-scale language models (LMs) requires huge amounts of text corpora. LMs for English enjoy ever growing corpora of diverse language resources. However, less resourced languages and their mono- and multilingual LMs often struggle to obtain bigger datasets. A typical approach in this case implies using machine translation of English corpora to a target language. In this work, we study the caveats of applying directly translated corpora for fine-tuning LMs for downstream natural language processing tasks and demonstrate that careful curation along with post-processing lead to improved performance and overall LMs robustness. In the empirical evaluation, we perform a comparison of directly translated against curated Spanish SQuAD datasets on both user and system levels. Further experimental results on XQuAD and MLQA downstream transfer-learning question answering tasks show that presumably multilingual LMs exhibit more resilience to machine translation artifacts in terms of the exact match score.

1 INTRODUCTION

Numerous research studies demonstrate how important the data quality is to the outcomes of neural networks and how severely they are affected by low quality data [13].

However, recently transfer learning, where a model is first pre-trained on a data-rich task before being fine-tuned on a downstream task, has emerged as a powerful technique in natural language processing. Models like T5 [11] are now showing human-level performance on most of the well-established benchmarks available for natural language understanding.

Yet, language understanding is not solved, even in well-studied languages like English, even when tremendous resources are used. This paper focuses on a less resourced language, Spanish, and pursues two goals.

First, we seek to demonstrate that data quality is an important component for training neural networks and overall increases natural language understanding capabilities. Hence, the quality of data should be considered carefully. We consider two recent Neural Machine Translated (NMT) SQuAD datasets, discussed in more detail in Section 4, for machine reading comprehension (MRC) in Spanish of different quality.

Second, after providing evidence of the data quality difference, we fine-tune both datasets on pre-trained multilingual and monolingual Spanish BERT models and see that there is a significant performance gap in terms of Exact Match (EM) and F1 scores in dev sets of the SQuAD family. However, unexpectedly, the results become more comparable when testing on external benchmarks recently proposed for cross-lingual Extractive QA.

Hence, we tackle the effects of both data quality and neural networks characteristics in an effort to demonstrate that both mentioned above are major factors in the outcomes and should be given equal respect in their primary design.

[1] SemanticWebCompany, Austria, maria.khvalchik@semantic-web.com
[2] TU Dresden & Fraunhofer IAIS, Germany, mikhail.galkin@tu-dresden.de

2 RELATED WORK

Historically, most of NLP tasks, datasets, and benchmarks were created in English, e.g., the Penn Treebank [10], SQuAD [12], GLUE [14]. Therefore, most of the large-scale pre-trained models were trained in the English-only mode, e.g., BERT [6] employed Wikipedia and the BookCorpus [17] as training datasets. Later on, the NLP community sought after increasing language diversity and multilingual models started to appear, such as mBERT or XLM [4].

However, large and diverse enough pre-training corpora of high quality often do not exist. Several methods have been developed to bridge this gap, e.g., applying machine translation frameworks to English corpora, or performing cross-lingual transfer learning [16]. Multilingual language models and pre-trained non-English language models are definitely in the focus of the NLP community. Still, the language understanding capabilities (hence, the performance) of language models largely depend on data collection and cleaning steps. In the MRC dimension, for instance, Italian SQuAD [5] is obtained via direct translation from the English version whereas French FQuAD [7] and Russian SberQuAD [8] have been created based on their language-specific part of Wikipedia often being much smaller than original SQuAD.

With the surge of language-specific pre-trained LMs several benchmarks have been developed that aim at evaluating multi- and cross-lingual characteristics of such LMs. Specifically, for the machine reading comprehension and question answering (QA) task there exist XQuAD [1] and MLQA [9]. In this work, we study how LMs perform in QA tasks in Spanish when fine-tuning on datasets of possibly different quality, i.e., directly machine translated and curated with the human-in-the-loop strategy.

3 PROBLEM FORMULATION

In this work, we aim at exploring the impact of machine translated corpora quality on downstream MRC tasks which is of high importance in less resourced languages. We consider Spanish as the target language as one of the most spoken languages in the world that nevertheless has a relatively little amount of available corpora for pre-training modern LMs for language understanding tasks. Taking this into account, we tackle the following research questions:

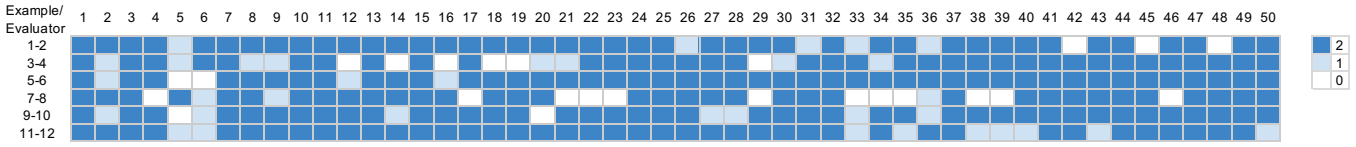

Figure 1. TAR translation evaluation heat map on 50 parallel SQuAD examples scored by 12 evaluators.

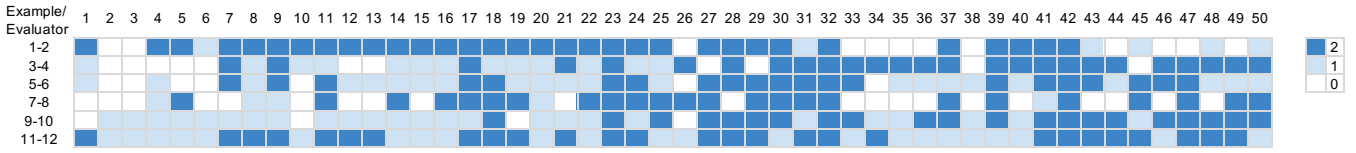

Figure 2. MT translation evaluation heat map on 50 parallel SQuAD examples scored by 12 evaluators.

- **RQ1:** Is there a quantifiable difference in data quality between machine translated and manually post-processed corpora for Spanish SQuAD datasets? In order to answer this question, we conduct a user study in Section 4.
- **RQ2:** Can we expect the performance difference of LMs in MRC QA tasks when fine-tuned on datasets of different quality? We perform an experimental study in Section 5.1.
- **RQ3:** Is there a performance difference in downstream transfer-learning QA tasks, i.e., on external benchmarks the LMs were not fine-tuned on? Experimental results are shown in Section 5.2.

4 USER STUDY

4.1 Data Sources

For the user study we employ the two following recent MRC Spanish translations:

TAR: prepared following the Translate-Align-Retrieve methodology which implies a lot of post-processing to improve the translation quality [2]. TAR SQuAD is produced from original English SQuAD corpus and contains both 1.1 and 2.0 versions. Further, each version contains datasets of two sizes, i.e., *regular* (or default) and *small* (half the size of the regular) that is less noisy and more refined.

MT: SQuAD 1.1 and 2.0 versions translated by a private European NMT company. [3]

4.2 Translation Evaluation

To estimate the quality of translation, 50 parallel examples from translated SQuAD 1.1 dataset were selected randomly. Twelve Spanish speaking evaluators were asked to give the following grades to 25 parallel examples each:

- 2, if understandable and there are only minor mistakes;
- 1, if understandable and has a few major mistakes;
- 0, if not understandable and has more than a few major mistakes.

In Table 1 the average translation evaluation score is depicted, from which we conclude that TAR translation is significantly better than MT translation.

To inspect further, in Figure 3 we provide the histogram of score frequencies where TAR translation produces 80% of the best scores while MT's produces only around 50%.

Table 1. Translation evaluation average.

Translation by	Average
TAR	1.717
MT	1.320

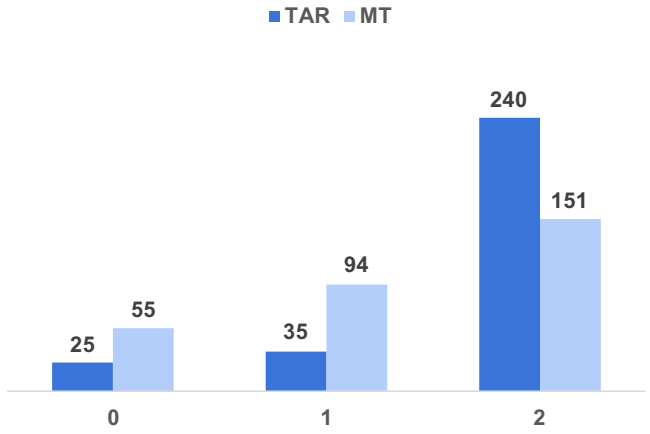

Figure 3. Score frequencies for TAR and MT translations.

Furthermore, to evaluate the agreement among raters, we aggregate the evaluations in heat map representation in Figure 1 and Figure 2. We can observe that most of the evaluators not only favoured but also synchronized well on the TAR scores, whereas MT scores appear to be more contrasting. This could possibly mean that the MT errors were so diverse that evaluators found the provided scale to some degree misleading. Hence, it was difficult to strictly represent the difference between minor and major errors.

Translation errors which could significantly affect the results have been collected and some examples are depicted in Table 4. The error types are the following: wrong gender inference, inaccurate translation or capitalization in named entities, adjectives misplacement regarding the noun. Here, we would like to point at the following errors in MT translation:

- an example of combining named entity's *"Warner Brothers"* in a literal Spanish translation as well as in the original state: *"... y hermanos Warner. Universal, Warner Brothers ..."*;

[3] We will publish the dataset on GitHub upon acceptance.

Table 2. Performance on es-SQuAD. All models are in the case-sensitive mode. Best column results in **bold**, second best underlined.

Model	es-SQuAD (TAR)								es-SQuAD (MT)			
	1.1				2.0				1.1		2.0	
	Small		Default		Small		Default		Default		Default	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
mBERT (1.1 Sm TAR)	**57.45**	73.34	55.04	72.38	-	-	-	-	56.24	71.12	-	-
mBERT (1.1 Def TAR)	56.30	73.71	59.65	76.32	-	-	-	-	54.80	71.67	-	-
mBERT (2.0 Sm TAR)	57.36	73.52	55.20	72.56	**59.85**	66.30	60.18	66.94	55.36	70.64	46.17	57.87
mBERT (2.0 Def TAR)	56.24	73.11	59.05	75.48	59.76	**67.17**	**62.08**	**68.90**	54.47	71.06	46.97	59.47
mBERT (1.1 MT)	54.25	71.42	53.90	71.90	-	-	-	-	61.20	64.19	-	-
mBERT (2.0 MT)	52.92	70.51	53.03	71.25	29.02	38.60	26.32	35.80	**61.27**	**74.15**	**61.46**	**74.72**
BETO	56.72	**74.38**	**59.71**	**76.92**	58.55	67.16	60.01	67.97	55.93	72.56	49.49	62.88
DistilledBETO	54.11	72.41	57.32	74.94	57.26	66.28	58.58	66.75	52.11	69.86	48.28	63.58

- an example of translation *"Universal Pictures"* by changing into plural form and dropping the noun *"universales"*;
- an example of *"the US War Department"* translation as an *"al Departamento de Guerra DE NOSOTROS"*, an impressive translation of a capitalized abbreviation of the United States.

Therefore, we can positively answer **RQ 1** as there indeed exists a substantial difference in corpora quality when applying additional post-processing over direct machine translated data.

5 EXPERIMENTAL STUDY

In the experimental study we evaluate the performance of Spanish LMs in machine reading comprehension tasks fine-tuning them on language corpora obtained via machine translation and translation with further rule-based post-processing.

Datasets. For fine-tuning the pre-trained LMs in Spanish we leverage different versions of es-SQuAD, i.e., TAR and MT described in Section 4. Small and Default versions of es-SQuAD (TAR) are annotated as *sm* and *def*, respectively. For benchmarking we employ dev sets of es-SQuAD datasets as well as test sets of MLQA [9] and XQuAD [1] where both context and question are in Spanish.

Models. We choose the pre-trained *mBERT-base-cased* for fine-tuning on es-SQuAD datasets. For a broader comparison we also employ already pre-trained and fine-tuned on SQuAD 2.0 Spanish-only LMs BETO [3] and its distilled version DistilledBETO.

Fine-Tuning Setup. The models are trained and evaluated in the cased mode using the HuggingFace Transformers [15] framework. When fine-tuning the default hyperparameters are used: three epochs of the Adam optimizer with an initial learning rate of 0.00005. The experiments are conducted on the Ubuntu 16.04 server equipped with one GTX 1080 Ti GPU and 256 GB RAM.

Metrics. We measure EM and F1 scores in each experiment as reported by task-specific evaluation scripts. For consistency reasons, we do not evaluate models fine-tuned on SQuAD 1.1 datasets against SQuAD 2.0 versions.

5.1 SQuAD Performance

In the first experiment, we fine-tune mBERT on TAR and MT datasets and evaluate their accuracy on the dev test of the respective tasks. That is, in order to study the impact of the fine-tuning dataset we optimize the model on TAR, but evaluate on the MT dev set, and vice versa. The empirical results are shown in Table 2.

First, we observe that LMs fine-tuned on the MT SQuAD considerably outperform other models in terms of both EM and F1 only on the MT dev set while being significantly inferior to all other models on the TAR dev test. For instance, mBERT fine-tuned on the MT-version of SQuAD 2.0 is about 12 EM and F1 points better than BETO on the MT-version of SQuAD 2.0 and at the same time is about 32 EM and F1 points worse than BETO in the default TAR-version of SQuAD 2.0.

Similarly, mBERT trained on the MT-version of SQuAD 1.1 achieves very good EM score on the MT-version dev set of SQuAD 1.1 but performs poorly on the TAR versions. Considering the difference in datasets quality demonstrated in Section 4, we deem that such a behavior is a sign of LMs sensitivity to artificially created corpora with numerous syntactic and semantic mistakes.

Moreover, the TAR-trained models show more consistent scores across the given tasks thus supporting the **RQ 2**, i.e., LMs tend to be more robust when trained and evaluated on well-prepared language corpora. Overall, in this experiment we find that LMs trained on Spanish-only corpora (e.g., BETO) perform on par or slightly better than massive multilingual LMs like mBERT fine-tuned on a similar task in a language-specific setting.

5.2 MLQA and XQuAD Performance

Table 3. Performance on MLQA and XQuAD. Cased models. Best in **bold**, second best underlined.

Model	MLQA		XQuAD	
	EM	F1	EM	F1
mBERT (1.1 Sm TAR)	42.74	64.36	53.61	72.89
mBERT (1.1 Def TAR)	43.14	66.44	54.62	75.30
mBERT (2.0 Sm TAR)	43.31	65.04	54.45	74.09
mBERT (2.0 Def TAR)	43.44	66.09	55.97	76.82
mBERT (1.1 MT)	44.43	64.83	**57.14**	75.46
mBERT (2.0 MT)	44.13	64.43	54.03	73.17
BETO	**45.12**	**68.77**	56.97	**78.15**
DistilledBETO	42.41	66.06	55.46	75.84

In the second experiment, we probe the TAR and MT fine-tuned models against MLQA and XQuAD in the Spanish context - Spanish question settings. The results are presented in Table 3. A clear winner is BETO which is pre-trained on Spanish-only corpora and

outperforms nearest contenders by about 2 F1 points. We then observe that TAR-trained models perform consistently better than MT-trained models in terms of F1 scores. Interestingly, in terms of EM scores mBERT 1.1 MT yields better performance than TAR and even language-specific models like BETO. Such a phenomena can be explained by robustness of large-scale multilingual LMs that might tend to generalize better over translation artifacts. We leave further research of this phenomena to the future work.

Overall, discussing the **RQ 3** we hypothesize that for downstream language-specific tasks LMs pre-trained in that specific language are more preferable. In case such a large-scale pre-training corpora is not available, well-processed machine translated sources tend to produce more robust LMs compared to purely machine translated sources.

6 CONCLUSION AND FUTURE WORK

In this work, we studied the impact of machine translated corpora quality on question answering tasks. Having formulated three research questions, we employed Spanish SQuAD-style datasets for empirical evaluation. User study confirmed there is a significant difference in dataset quality and amount of language artifacts. Further experimental studies confirmed that LMs are sensitive to the quality of machine translated corpora. We also observe signs of LMs robustness to translation defects in downstream transfer learning tasks.

For the future work we pose a question towards conducting an appropriate analysis on how neural networks overcome the flaws in the data, being not always machine translated, to become robust and noise resilient.

REFERENCES

[1] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama, 'On the cross-lingual transferability of monolingual representations', *CoRR*, (2019).

[2] Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa, 'Automatic spanish translation of the squad dataset for multilingual question answering', *CoRR*, (2019).

[3] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez, 'Spanish pre-trained bert model and evaluation data', in *to appear in PML4DC at ICLR 2020*, (2020).

[4] Alexis Conneau and Guillaume Lample, 'Cross-lingual language model pretraining', in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pp. 7057–7067, (2019).

[5] Danilo Croce, Alexandra Zelenanska, and Roberto Basili, 'Neural learning for question answering in italian', in *AI*IA 2018 – Advances in Artificial Intelligence*, eds., Chiara Ghidini, Bernardo Magnini, Andrea Passerini, and Paolo Traverso, pp. 389–402, (2018).

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, '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, NAACL-HLT 2019*, pp. 4171–4186, (2019).

[7] Martin d'Hoffschmidt, Maxime Vidal, Wacim Belblidia, and Tom Brendlé, 'Fquad: French question answering dataset', *CoRR*, (2020).

[8] Pavel Efimov, Leonid Boytsov, and Pavel Braslavski, 'Sberquad - russian reading comprehension dataset: Description and analysis', *CoRR*, (2019).

[9] Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk, 'MLQA: evaluating cross-lingual extractive question answering', *CoRR*, (2019).

[10] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz, 'Building a large annotated corpus of english: The penn treebank', *Computational Linguistics*, **19**(2), 313–330, (1993).

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, 'Exploring the limits of transfer learning with a unified text-to-text transformer', *CoRR*, (2019).

[12] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, 'Squad: 100, 000+ questions for machine comprehension of text', in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 2383–2392, (2016).

[13] Valerie Sessions and Marco Valtorta, 'The effects of data quality on machine learning algorithms', in *Proceedings of the 11th International Conference on Information Quality, MIT, Cambridge, MA, USA, November 10-12, 2006*, eds., John R. Talburt, Elizabeth M. Pierce, Ningning Wu, and Traci Campbell, pp. 485–498. MIT, (2006).

[14] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman, 'GLUE: A multi-task benchmark and analysis platform for natural language understanding', in *7th International Conference on Learning Representations, ICLR 2019*, (2019).

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew, 'Huggingface's transformers: State-of-the-art natural language processing', *ArXiv*, (2019).

[16] Shijie Wu and Mark Dredze, 'Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT', 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 2019*, pp. 833–844, (2019).

[17] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler, 'Aligning books and movies: Towards story-like visual explanations by watching movies and reading books', in *2015 IEEE International Conference on Computer Vision, ICCV*, pp. 19–27, (2015).

Table 4. A sample of TAR and MT translations error examples annotated in the user study.

English	TAR	MT
Gender inference (TAR avg score = 0.8, MT avg score = 0)		
It's not clear, however that this stereotypical view reflects the reality of East Asian classrooms or that the educational goals in these countries are commensurable with those in Western countries. In Japan, for example, although average attainment on standardized tests may exceed those in Western countries, classroom discipline and behavior is highly problematic. Although, officially, schools have extremely rigid codes of behavior, in practice many teachers find the students unmanageable and do not enforce discipline at all.	Sin embargo, no está claro que esta opinión estereotipada refleje la realidad de las aulas de Asia oriental o que los objetivos educativos de esos países sean acordes con los de los países occidentales. En Japón, por ejemplo, aunque el rendimiento medio de los ensayos estandarizados puede superar los de los países occidentales, la disciplina y el comportamiento de las aulas son muy problemáticos. Aunque oficialmente las escuelas tienen códigos de conducta extremadamente rígidos, en la práctica muchos maestros consideran que los estudiantes son inmanejables y no aplican la disciplina en absoluto.	No está claro, sin embargo, que esta visión estereotipada refleja la realidad de las aulas del Asia oriental o que los objetivos educativos en estos países son conmensurables con los de los países occidentales. En Japón, por ejemplo, aunque el logro promedio de las pruebas estandarizadas puede exceder las de los países occidentales, la disciplina y el comportamiento en el aula son altamente problemáticos. Aunque, oficialmente, las escuelas tienen códigos de comportamiento extremadamente rígidos, en la práctica muchos profesores encuentran a los estudiantes inmanejables y no aplican la disciplina en absoluto.
Translation and capitalization inconsistency in named entities (TAR avg score = 1, MT avg score = 0.1)		
The motion picture, television, and music industry is centered on the Los Angeles in southern California. Hollywood, a district within Los Angeles, is also a name associated with the motion picture industry. Headquartered in southern California are The Walt Disney Company (which also owns ABC), Sony Pictures, Universal, MGM, Paramount Pictures, 20th Century Fox, and Warner Brothers. Universal, Warner Brothers, and Sony also run major record companies as well.	La industria del cine, la televisión y la música se centra en Los Ángeles en el sur de California. Hollywood, un distrito dentro de Los Ángeles, es también un nombre asociado a la industria cinematográfica. Con sede en el sur de California están The Walt Disney Company (que también posee ABC), Sony Pictures, Universal, MGM, Paramount Pictures, 20th Century Fox, y Warner Brothers. Universal, Warner Brothers y Sony también tienen grandes compañías discográficas.	La imagen del movimiento, la televisión y la industria musical se centran en los Ángeles en el sur de California. Hollywood, un distrito de los Ángeles, es también un nombre asociado a la industria fotográfica de movimiento. Con sede en el sur de California están la compañía Walt Disney (que también posee ABC), imágenes de Sony, universales, MGM, imágenes principales, Fox del siglo 20 y hermanos Warner. Universal, Warner Brothers y Sony también dirigen grandes empresas de registro.
During the same year, Tesla wrote a treatise, The Art of Projecting Concentrated Non-dispersive Energy through the Natural Media, concerning charged particle beam weapons. Tesla published the document in an attempt to expound on the technical description of a "superweapon that would put an end to all war." <...> Tesla tried to interest the US War Department, the United Kingdom, the Soviet Union, and Yugoslavia in the device.	Durante el mismo año, escribió un tratado, The Art of Projecting Concentrated nondispersive Energy through the Natural Media, sobre las armas de haz de partículas cargadas. Tesla publicó el documento en un intento de exponer la descripción técnica de una "superarma que pondría fin a toda guerra". <...> Tesla trató de interesar al Departamento de Guerra de los Estados Unidos, el Reino Unido, la Unión Soviética y Yugoslavia en el dispositivo.	Durante el mismo año, Tesla escribió un treatise, el arte de proyectar energía concentrada no dispersa a través de los medios naturales, en relación con las armas de haz de partículas cargadas. Tesla publicó el documento en un intento de exponer la descripción técnica de un «superarma que pondría fin a toda guerra». <...> Tesla trató de interesar al Departamento de Guerra DE NOSOTROS, al Reino Unido, a la Unión Soviética y a Yugoslavia en el dispositivo.
Adjectives placement regarding the noun (TAR avg score = 1, MT avg score = 0)		
CBS broadcast Super Bowl 50 in the U.S., and charged an average of $5 million for a 30-second commercial during the game. The Super Bowl 50 halftime show was headlined by the British rock group Coldplay with special guest performers Beyoncé and Bruno Mars, who headlined the Super Bowl XLVII and Super Bowl XLVIII halftime shows, respectively. It was the third-most watched U.S. broadcast ever.	CBS transmitió el Super Bowl 50 en los Estados Unidos, y cobró un promedio de $5 millones por un comercial de 30 segundos durante el juego. El espectáculo de medio tiempo del Super Bowl 50 fue encabezado por el grupo de rock británico Coldplay con artistas invitados especiales como Beyoncé y Bruno Mars, quienes encabezaron los shows de medio tiempo del Super Bowl XLVII y Super Bowl XLVIII, respectivamente. Fue el tercer programa más visto de Estados Unidos.	CBS emitió 50 super bowl en los U. S. y cobró un promedio de US $5 millones por un 30 - segundo comercial durante el juego. El espectáculo de semáforo 50 de super fue encabezado por el grupo de rock británico Coldplay con artistas invitados especiales Beyoncé y Bruno Mars, que encabezaron el súper súper XLVII y los espectáculos de semestral XLVIII, respectivamente. Fue la tercera, la más observada u.