recibido 16-12-2022 revisado 31-01-2023 aceptado 07-02-2023

Widaug. Data augmentation for named entity recognition using Wikidata

Widaug. Aumento de datos para el reconocimiento de entidades nombradas usando Wikidata

Pablo Calleja, Alberto Sánchez, Oscar Corcho
Ontology Engineering Group
Universidad Politécnica de Madrid
{p.calleja,o.corcho}@upm.es

Abstract: The current state of the art of Natural Language Processing models are based on the use of a big amount of data to be trained. The more, the better. However, this is quite a limitation in the creation of datasets for specific natural language processing tasks such as Named Entity Recognition, which involves one or more annotators to read, understand and annotate those required named entities along a corpus. Currently, there are many good general domain corpora for the English language. However, particular domains or scenarios and other non-English languages are still not so represented in the research community. Thus, data augmentation techniques are explored to create synthetic data similar to the originals to enrich the training process of the models. On the other hand, knowledge graphs contain a lot of valuable information that is not being used to help in the data augmentation process. This work proposes a data augmentation method based on the Wikidata knowledge graph which is tested in a Spanish corpus for a Named Entity Recognition challenge.

Keywords: Named Entity Recognition, data augmentation, Wikidata.

Resumen: El estado del arte actual de los modelos de Procesamiento de Lenguaje Natural se basa en el uso de una gran cantidad de datos para ser entrenados. Cuantos más, mejor. Sin embargo, esto es una gran limitación en la creación de conjuntos de datos para tareas específicas de procesamiento de lenguaje natural, como el reconocimiento de entidades nombradas, que involucra a uno o más anotadores para leer, comprender y anotar las entidades nombradas requeridas a lo largo de un corpus. Actualmente, hay bastantes corpus buenos de dominio general para el inglés. Sin embargo, los dominios o escenarios particulares y otros idiomas distintos del inglés aún no están tan representados en la comunidad de investigación. Por ello, se exploran técnicas de aumento de datos para crear datos sintéticos similares a los originales para luego enriquecer el proceso de entrenamiento de los modelos. Por otro lado, los grafos de conocimiento contienen muchísima información valiosa que no se está utilizando para ayudar en el proceso de aumento de datos. Este trabajo propone un método de aumento de datos basado en el grafo de conocimiento de Wikidata que es evaluado en un corpus español para un desafío de reconocimiento de entidades nombradas.

Palabras clave: Reconocimiento de Entidades Nombradas, aumento de datos, Wikidata.

1 *Introduction*

Named Entity Recognition (NER) is a Natural Language Processing (NLP) task that consists of identifying named entities in the text and classifying them. Traditionally, those name entities are proper names of persons, locations, organizations or miscellaneous proper names (Grishman and Sundheim, 1996; Tjong Kim Sang, 2002). Nowadays, the original classification groups are extended and adapted to particular domains or to scenarios of interest. For instance, the biomedical domain has defined its own relevant classification groups such as diseases, chemical compounds, DNA, etc. (Perera,

ISSN 1135-5948 DOI 10.26342/2023-70-12

Dehmer, and Emmert-Streib, 2020; Asghari, Sierra-Sosa, and Elmaghraby, 2022). Moreover, recent challenges and corpora propose to identify new classification groups, such as drugs (Li, Zhang, and Zhou, 2020), complex named entities (Malmasi et al., 2022) or software mentions (Schindler et al., 2021). Although the classification groups are highly different between them and the instances could not be proper names, the task is still based on the identification of particular terms or nouns that belong to a particular classification group.

Currently, the best results obtained in Named Entity Recognition are usually based on pretrained language models[1] which have been fine-tuned for the task. Although the finetuning process requires less data to achieve good results, models still need a training corpora as large as possible to learn from it. However, the creation of annotated corpora is a process that requires a huge human effort that involves reading, understanding, and annotating particular entities in the text. Also, more than one annotator per document is usually required to achieve a good quality dataset, in which the inter-annotation agreement is validated.

The lack of data in certain domains makes it impossible to create efficient models for a NER task. An example of this is the case of the SmarTerp project (Rodriguez et al., 2021), a project to help interpreters in simultaneous interpretation contexts in the European Parliament. This project required the implementation of a named entity recognition system for 'important' words within a specific scope. Entities such as locations or dates are required but also specific terms about the topic being discussed, such as 'types of soil for cultivation' or 'types of metal forging procedures'. This type of 'important' named entities were specific for each European Parliament session related to the topic which was discussed and in different European languages. However, due to the ambiguity of the problem and the lack of prepared annotated corpora, the development of a specific annotated corpus was one of the most difficult tasks.

Beside the problem of lack of data in certain scenarios, the language problem is also added. Other non-English languages are underrepresented in terms of corpora and models. A common approach to solve this gap is the proposal of challenges and shared tasks for different languages. For instance, in the last years the Spanish community have released datasets for challenges in particular domains. Two of the most important ones have been LivingNER challenge (Farré-Maduell et al., 2022) which is focused on the identification of living entities providing a corpus of 2000 annotated clinical cases for training, development and testing, and CANTEMIST (Miranda-Escalada, Farré, and Krallinger, 2020), which is focused on the identification of tumor morphology providing a corpus of 1301 annotated oncological clinical cases for the same three tasks. Both challenges have been an important improvement on the Spanish language community, but they are still small compared to others that English language models use in other scenarios.

Data augmentation is a technique that is used to generate new synthetic data based on the modification of the original data in those cases where there are not enough samples to train a machine learning model and to achieve better results. This technique has been used and adapted in different research areas. In the NLP context, data augmentation is usually based on adding noise (removing/adding words) to original sentences, adding synonyms or moving words to other positions (Erd et al., 2022; Dai and Adel, 2020). However, there are not so many works that exploit knowledge graphs to acquire structured data to enrich the data augmentation process.

The objective of the work consists in the creation of a method named Widaug using information extracted from the common well-known Wikidata knowledge graph. The target of this method is to cover scenarios such as the Smarterp project in which just a few sentences could be annotated by experts, making it impossible to create a representative corpus. Our hypothesis is that the proposed data augmentation method can improve the performance of the training model better than other traditional techniques, specifically for small corpora, relying on the knowledge provided by Wikidata.

For the evaluation of the method, a set of experiments have been performed on the second task of ProfNER's challenge (Miranda-Escalada et al., 2021) which proposes a na-

[1] https://paperswithcode.com/sota/token-classification-on-conll2002

med entity recognition problem for the recognition of 'professions' within tweets related to the COVID-19 pandemic. The code and the experiments can be found in our public GitHub repository.[2]

The paper is structured as follows: Section 2 details the state-of-the-art of data augmentation and Section 3 presents the approach of the method, the use case and the experiments performed. Section 4 evaluates the results and, finally, Section 5 achieves the conclusions and future work.

2 State of the art

Data augmentation has been a widely research area that has been involved in many different tasks in which machine learning models have significance such as computer vision or, in this particular, for NLP tasks. The basic idea is to take partial data or related data as seeds to create a larger dataset to train a machine learning model. With more data, the model will be able to generalize better and obtain better results.

In NLP there are some common techniques such as replacement of words. Replacement is one of the first techniques used in data augmentation. External resources such as Wordnet (Zhang, Zhao, and LeCun, 2015) or Word2vec (Wang and Yang, 2015) are used for synonym replacement. Recent approaches (Wu et al., 2019) used pretrained language models to replace words that are suitable in the position of the word in the original sentence, by doing masks. Other works do mention replacement; in the training data the entities are replaced for others from a manually created dictionary that contained entities that were not part of the training data (Liu et al., 2020).

The Easy data augmentation techniques (EDA) were presented for text classification (Wei and Zou, 2019), which are based on synonym replacement, random insertion, random swap and random deletion. Other work extended EDA techniques for NER tasks using the UMLS knowledge base (Kang et al., 2021). Currently, there are libraries such as NLPAug[3] or TextAugment (Marivate and Sefara, 2020) that facilitate the implementation process for most of these techniques.

Back translation is also a common technique which is based on the retranslation of content from a target language back to its source language. The purpose in data augmentation is to get similar sentences with changes in some words that have been modified in the translation process. It can be used for topic classification (Xie et al., 2020) or sentiment analysis classification (Luque, 2019). Additionally, this approach has been tested for NER tasks (Yaseen and Langer, 2021). In the Spanish language, there are works that have presented back translation techniques for text classification problems (Luo, 2021; Guzman-Silverio, Balderas-Paredes, and López-Monroy, 2020)

In addition, text generation or sentence generation is a technique in which new synthetic sentences are created using language models or generative models to extend the original data. There are works that use this technique on text classification tasks (Bayer et al., 2022) or NER tasks (Ding et al., 2020).

Related to knowledge graphs, works that have explored the use of Wikidata (Raiman and Miller, 2017) have proposed a general scheme to do mention replacement for the general types (person, location, dates, etc.) for a Question Answering task using the instances of the general types represented in Wikidata. Other works (Kim, Kim, and Kang, 2022) used Wikidata to extract aliases of named entities with the label *Also known as* which are used for mention replacement.

The method proposed in this work combines the sentence generation approach using the information represented in the Wikidata concepts, using its labels and the most important relations to create new sentences in combination with sentences extracted from the Wikipedia pages of the concepts. Moreover, back translation and mention replacement approaches are tested.

3 Methodology

This section presents the use case in which the data augmentation method has been tested, the proposed data augmentation method using Wikidata, the experiments design and how the models have been fine-tuned with the augmented data.

3.1 Use case

The method for data augmentation has been tested for the ProfNER challenge (Miranda-Escalada et al., 2021). This challenge is part of the Social Media Mining for Health

[2] https://github.com/oeg-upm/widaug

[3] https://github.com/makcedward/nlpaug

(SMM4H), an initiative that seeks the application of machine learning methods for the extraction of information in social networks and its use in the health sector. The ProfNER-ST challenge, in particular, seeks to identify professions and occupations on social networks in Spanish within the healthcare field.

This challenge is particularly relevant in the context of the COVID-19 pandemic and its repercussions on mental health. Therefore, the annotated data were obtained using a web crawler on Twitter using keywords such as 'Covid-19', 'epidemic' or 'confinement'. The main aim is to use Natural Language Processing to identify vulnerable occupations in this context. This vulnerability can be both direct (health professionals in the first line of contact) and indirect (professions such as drivers, guards, carers, etc.).

Specifically, the use case has focused on the second task of the challenge, which seeks the identification and classification of professions in tweets related to the COVID-19 pandemic. The challenge provides the training, development and test corpus sets. The train set contains around 6,000 annotated tweets and the validation set contains around 2,000 annotated tweets.[4] As the gold annotations from the test set are not released, the development set is used for the evaluation of the data augmentation method.

The types of the named entities are: *PROFESION* (profession) which are entities referred to a profession that provides a salary such as 'doctor' or 'driver', *SITUACION_LABORAL* (employment situation) which are entities referred to an specific working condition such as 'worker' or 'self-employed', *ACTIVIDAD* (activity) which are unpaid works such as 'volunteer' and *FIGURATIVA* (figurative) which are used to mention metaphoric works such as 'joker' or 'pseudo journalist', usually used as sarcasm or jokes.

In total, there are 2,597 entities classified as PROFESSION (2,163), WORKING SITUATION (349), ACTIVITY (61) and FIGURATIVE (24). This work has focused only on those entities of type 'PROFESSION', which is the most representative named entity type in the corpus. The other types have not been considered by their ambiguity and their under-representation in the corpus. Of the 2,161 profession named entities, 1,532 are single words such as 'president' and 631 are multiword terms that refer to a profession such as 'bus driver' or 'national policeman'.

Moreover, the training and development corpus have been cleaned in order to avoid emojis, urls, hashtags and other characters outside of the scope of the utf-8 chars recognized by the tokenizer of the language model. Also, sentences with less than four tokens without any named entity annotated are removed.

3.2 Proposed method

The general approach of the Widaug method is presented in Figure 1. The method needs a corpus with annotated named entities, the language of the corpus and the target type of the named entities that will be augmented. The method performs the following tasks.

First, the method extracts the entities from the corpus that belong to the target type. Then, the method queries Wikidata to obtain instances of the target type. This search is performed by querying for instances of the concept in the graph that corresponds to the target type. The relation of the instance is represented by the relation '*instanceOF*' (wdt:P31). The instances are stored and tagged as candidate named entities.

The next step is to perform a filter in order to obtain only those named entity candidates close to the domain of the original annotated named entities extracted in the first step. Wikidata represents a huge amount of information that even the instances of a similar class could not have the same semantic meaning for the target domain of the corpus. The filter is carried out with a word embedding method, in which a candidate has to be up to 70 % similar, using the cosine similarity, for more than one original named entity. Capturing entities with more than one 70 % similarity can represent a trend or partial topic for similar named entities and avoids outliers. This value has been considered based on previous studies (Rekabsaz, Lupu, and Hanbury, 2017) and a preliminary study in which clearly unwanted terms over 50 % of similarity (not actual works such as prefect of Rome or controversial works such as sexual works) are analyzed to be below the selected threshold.

As word embeddings pretrained models do

[4] https://zenodo.org/record/4563995#.Y5cuPuzML0o

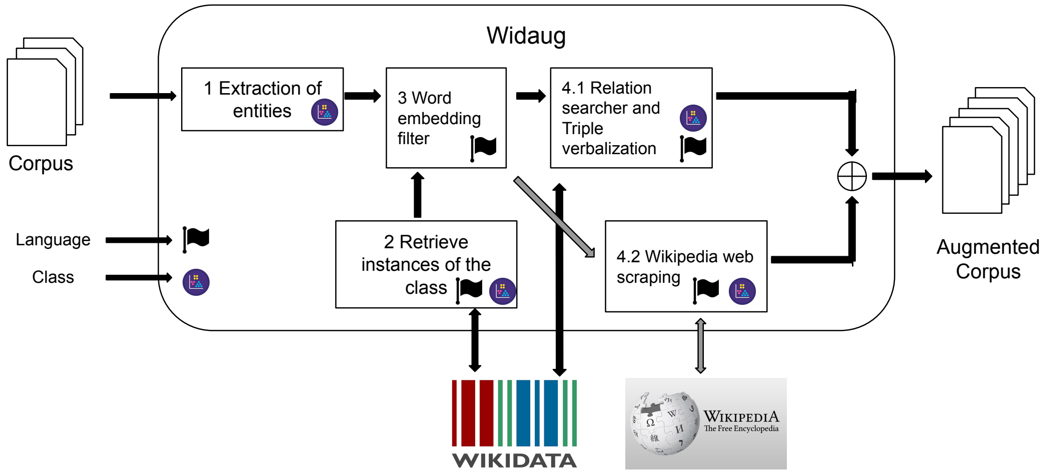
Widaug
Corpus
Language
Class
1 Extraction of entities
3 Word embedding filter
4.1 Relation searcher and Triple verbalization
2 Retrieve instances of the class
4.2 Wikipedia web scraping
Augmented Corpus
WIKIDATA
WIKIPEDIA The Free Encyclopedia

Figura 1: Overview of the data augmentation method Widaug.

not represent multi-word terms, the embedding filter is performed with the principal noun of the term of the candidate over the rest of principal nouns of the original named entities.

Once the candidates have been filtered, the proposed method exploits the information from the Wikidata concept, based on its properties and relations, and verbalizes it. The property to be found is *schema:description*, which provides a short description of the concept in natural language, and the following relations: P279 (subclass_of), P1056 (produce), P2283 (use) and P425 (field of occupation). The description property and the related concepts require to have their label property in the target language to be extracted. For instance, if a concept contains descriptions but not in the required language, the property is not extracted. The selection of these three relations apart from subclass_of is based on a previous study to find the three most repeated relations with labels in the required language in all the candidates. This method could be extended for other named entity types.

Then the properties and relations are verbalized. An example of this is shown in Table 1 with the Wikidata concept of baker for the English language. For the description, the verb 'to be' is used to join the named entity with its description. The named entity is annotated with the tag of the target type. In the rest of the relations, more than one element can be represented. For instance, in the example of the relation 'use' for baker, three elements are retrieved (heat, oven and bakery). The verbalization is used using the same verbs that represent the relation and the concept is also annotated. In the case of subclass of, the elements retrieved are also tagged with the same type (because they are subclasses). At this moment, the developed method covers the verbalization of sentences for the English and Spanish languages.

Finally, the last task consists of generating sentences based on web scrapping. This approach consists in capturing sentences from the Wikipedia page of each candidate (as most of the concepts have one) in order to create well-structured natural language sentences. Only sentences that contain the named entity are captured. Finally, those sentences are tokenized and the named entity is labeled according to its classification.

3.3 Experiment design

For the evaluation of the method, different experiments have been performed. First, the original training corpus has been randomly sampled at 10, 30, 50 and 100 %. The idea is to evaluate the performance of the method with different subsets of the original data and with the complete corpus such as Erd's evaluation (Erd et al., 2022). Then, four data augmentation methods are performed over the prepared corpora: the proposed augmentation method based on Wikidata (Widaug), a mention replacement method, a back translation method and finally, a combination of Widaug and back translation. All the augmented corpora are used to fine-tune a Spanish language model for the Named Entity Recognition task (a.k.a. token classification

	Examples								
Description	baker	is	the	persons	who	prepares	or	sells	bread
	B	O	O	O	O	O	O	O	O
Subclass of	baker	is	a	type	of	artisan	-	-	-
	B	O	O	O	O	B	-	-	-
Produce	baker	produces	bread	-	-	-	-	-	-
	B	O	O	-	-	-	-	-	-
Use	baker	uses	heat	oven	and	bakery	-	-	-
	B	O	O	O	O	O	-	-	-
Field of occupation	baking	is	the	field	of	occupation	of	baker	-
	O	O	O	O	O	O	O	B	-

Tabla 1: Example of verbalization of relations of the Wikidata concept ‘baker’ in English language. Tag ‘B’ represents the label B-Profession.

task). Moreover, the corpus without augmentation is used as a base line. The target of the experiments is to show how the methods perform augmenting the data for training and to check if it is possible to reach higher performance rather than without them. Figure 2 shows the overview of the experiments.

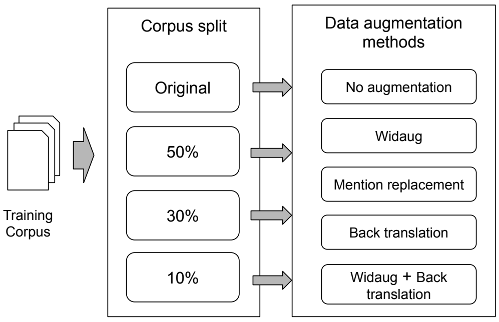

Figura 2: Experiments performed.

The proposed Widaug method has been adapted as follows. First, the target language needed is Spanish. Then, the target classification group is ‘profession’ which is represented in Wikidata as the concept wd:Q28640, which has 2,720 instances. The filtering process is performed with a FastText Spanish word embedding model (Bojanowski et al., 2016). After the filtering process, the candidate named entities are reduced to 852. The process of augmenting data from Wikidata generated 532 new annotated sentences and the generation of sentences from Wikipedia 855 new sentences.

The mention replacement method has been configured as follows. The 852 previously filtered candidates have been used to replace the mentions (named entities) in the corpus to generate new sentences, based on the work of Jonathan Raiman (Raiman and Miller, 2017). All the sampled corpus (10 %, 30 % and 50 %) have been augmented until the size of the original corpus (100 %). For the 100 % corpus, it has been augmented a 50 % more.

The back translation method is based on the library *BackTranslation*[5]. For each sentence of the corpus that contains an annotated named entity, a new sentence is generated translating it to English language and back to Spanish language. Finally, a combination of the Widaug method and the back translation method is proposed to evaluate the performance of one of the most representative methods for Spanish language in combination with the proposed one.

3.4 Fine-tuning process

For the fine-tuning process, the *Google Collaboratory* platform has been used to use GPUs for acceleration. Usually, a Tesla T4 GPU is given to train the models. The different trainings have been carried out for six epochs, which has been seen to be the point in which the original training corpus does not improve more.

The language model used is the MarIA (Gutiérrez-Fandiño et al., 2022) model developed at the Barcelona Supercomputing Center (BSC) with the database of the *Biblioteca Nacional Española* (National Library of Spain). Currently, MarIA models are the

[5] https://pypi.org/project/BackTranslation/

best models in terms of performance publicly available for Spanish language. For instance, the Rigoberta model (Serrano et al., 2022) claims to outperform MarIA results, but is not public.

This model is based on the RoBERTa architecture and the dataset contains 570 GB of cleaned training data. Although there are several different models, the model to use would be the *base*[6] model, which has 12 layers, 768 hidden layers and 125M parameters.

Hyperparameters of the training model have been 16 of batch size, 500 warm-up steps, 0.01 of weight decay and 1e-4 of learning rate and the models have been trained with the Huggingface transformers library. All models are evaluated with the validation corpus of the challenge.

4 Evaluation

This section details the evaluation results obtained from the experiments. As a traditional Named Entity Recognition task or token classification problem, the metrics used in the evaluation are precision, recall and f-measure. In addition, a discussion and an analysis error are performed.

4.1 Obtained results

Table 2 shows the results obtained for each portion of the original corpus (10, 30, 50 and 100 %) and for each data augmentation approach (mention replacement, back translation, Widaug and the combination of Widaug and back translation). Additionally, the results of the training corpus without data augmentation are presented as a baseline. The results presented for each combination of corpus and approach is the best result obtained within the 6 epochs of training.

The results obtained for the not augmented data (No Aug) for all created corpus (10, 30 and 50 %) have the lower values of all experiments in terms of the F-measure. However, none of the data augmentation methods have improved the results obtained for the original training corpus. Data augmentation methods have added noise to the training process.

In contrast, the data augmentation methods (Mention Replacement (MR), Back Translation (BT) and Widaug) have improved all the results over the baseline for all corpus, being Widaug the best performance method over the rest, followed by back translation. However, the final experiment in which Widaug is combined with back translation (BT) has not improved the overall results; only in the 30 % corpus is slightly improved (0.02).

4.2 Discussion and error analysis

The results obtained confirm the hypotheses of the work, that the proposed data augmentation method can improve the performance of the training model better than other traditional methods, specifically for small corpora such as the 10 % corpus which has an improvement of 0.11 in the F-measure. However, an error analysis has been performed and studied to understand the behaviour of the trained models.

First, the corpus presents some limitations. The size of the original training corpus is quite small (12,707 sentences), which has been reduced in the cleaning process to 11,050 sentences, with entities tended to repeat themselves a lot such as *sanitario* (sanitary) and *guardia civil* (civil guard). Also, the corpus is comprised of sentences of tweets, which contains typos and unstructured information (e.g., several mentions to public charges and services to advise them). Therefore, the training process with a generic language model with this corpus has a limitation.

Moreover, the validation corpus, which has been used for testing, has named entities out of the scope of common knowledge and that are not present either in the training corpus. For instance, the named entity 'tcae' is annotated and means *Técnico en Cuidados Auxiliares de Enfermería* (Nursing Auxiliary Care Technician). All the models fail in the recognition of these kinds of named entity and the results are never improved more than the 78 % of F-Measure.

Also, it is important to highlight that the mention replacement method has not achieved the expected results presented in other works of the state-of-the-art. Analyzing the context of the use case, we have consider than mention replacement needs from more representative natural language sentences with named entities to be switched. The original corpus is comprised of short tweets that contain profession named entities, sometimes, without any context. On the contrary, the back translation method has achieved good

[6] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne

	10 %			**30 %**			**50 %**			**100 %**		
	p	r	F	p	r	F	p	r	F	p	r	F
No Aug	0.792	0.471	0.591	0.761	0.685	0.721	0.782	0.695	0.736	0.829	0.746	**0.785**
MR	0.640	0.685	0.661	0.732	0.692	0.711	0.752	0.728	0.740	0.752	0.728	0.740
BT	0.738	0.654	0.693	0.739	0.709	0.724	0.775	0.690	0.730	0.788	0.756	0.771
Widaug	0.756	0.667	**0.709**	0.833	0.647	**0.728**	0.810	0.695	**0.748**	0.819	0.718	0.765
Widaug+BT	0.733	0.670	0.700	0.787	0.698	**0.740**	0.787	0.698	0.740	0.797	0.748	0.772

Tabla 2: Evaluation results measuring Precision (p), Recall (r) and F-measure (F) for the four corpus (10, 30, 50 and 100 %). Each row corresponds to a data augmentation method: no augmentation (No Aug), mention replacement (MR), back translation (BR), the proposed method (Widaug) and the combination of Widaug and back translation (Widaug+BT).

results for two main reasons. The first one is that some named entities have changed their gender in the process of translating to English language. Words that in Spanish were feminine, come back as masculine, doing a good data augmentation process. The second one is that some named entities of the corpus have changed their original language; there are cases of Catalan mentions that are back translated to Spanish or cases in which the English terms are also accepted in Spanish language such as *animadora* (cheerleader). Even though back translation have achieve close results to Widaug, we cannot generalize that could be similar for other scenarios without the properties of this particular corpus.

5 Conclusions and future lines

This paper has shown a simple data augmentation approach based on the use of Wikidata as a source of information. Knowledge graphs represent and link concepts and information already validated by humans, and this type of resource has not been exploited at all in the generation of new synthetic data for data augmentation.

The results show that there is a significant improvement for small datasets. For instance, the improvement of the 10 % corpus has been up to 0.11 more. Therefore, this method will cover scenarios in which it is difficult to find annotated data without involving human effort. Moreover, one of the benefits of using Wikidata and Wikipedia as an external resource to generate new data is that the new sentences are still human-readable and do not contain language errors, as approaches such as synonym replacement or random deletion may produce.

However, the method does not reflect a significant improvement over the full dataset. In these particular experiments we have discovered, as the discussion section presents, that the difficulties presented in the validation corpus make difficult to achieve better results over the training with the original corpus.

Analyzing the results and the discussion lines, the method has different future lines which should be explored. Some of them are:

- The use of a generic knowledge graph such as Wikidata does not allow the application of this method to more specific scenarios. For instance, for the identification of diseases within a biomedical domain. Wikidata does not contain specific domain information such as MESH or SNOMED-CT. Therefore, the use of knowledge graphs of other specific domains could be explored. This could allow for a higher level of detail in the generation of new synthetic data, which could lead to better quality results.
- Wikidata contains a huge amount of heterogeneous information. This is why many of the entities extracted, although correct, are far from the target domain and should be filtered. An example of this can be seen in the context of the 'profession' itself, where some of the retrieved entities were far from the context of the use case (for instance: 'chess referee', 'esperantologist', 'primatologist', etc). The word embedding filter is a key process to get better results and not add noise to the training corpus. However, pretrained word embedding models do not represent n-gram terms. Thus, pro-

fessions with n-gram terms are not being filtered correctly (e.g., *guardia civil* (civil guard)). So, new approaches will focus on the identification of the correct vectors for those terms, using, for instance, sentence embeddings which are based on language models.

- Moreover, Wikidata has a gender bias. Most of the concepts are presented with male gender and the new synthetic data are created in the same way. The next step is to identify the impact of generating instances in the female gender.
- Synthetic data generation with language adaptation of the original sentences. In this particular use case, we have found that the syntactic structures of natural language sentences generated or extracted are far different from the original ones that are tweets. Thus, a better adaptation to the original style, in terms of terminology and syntactic structure, should be done.

Acknowledgements

Financed by the European Union-NextGenerationEU (UP2021-035), by the SmarTerp Project (EIT-Digital-21184) and by the project HCommonK (RTC2019-007134-7, funded by MCIN/AEI/ 10.13039/501100011033).

References

Asghari, M., D. Sierra-Sosa, and A. S. Elmaghraby. 2022. Biner: A low-cost biomedical named entity recognition. *Information Sciences*, 602:184–200.

Bayer, M., M.-A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, and C. Reuter. 2022. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. *International Journal of Machine Learning and Cybernetics*, pages 1–16.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2016. Enriching word vectors with subword information.

Dai, X. and H. Adel. 2020. An analysis of simple data augmentation for named entity recognition. *arXiv preprint arXiv:2010.11683.*

Ding, B., L. Liu, L. Bing, C. Kruengkrai, T. H. Nguyen, S. Joty, L. Si, and C. Miao. 2020. Daga: Data augmentation with a generation approach for low-resource tagging tasks. *arXiv preprint arXiv:2011.01549.*

Erd, R., L. Feddoul, C. Lachenmaier, and M. J. Mauch. 2022. Evaluation of data augmentation for named entity recognition in the german legal domain. In *AI4LEGAL-KGSUM 2022 Artificial Intelligence Technologies for Legal Documents and Knowledge Graph Summarization 2022*, number 3257 in CEUR Workshop Proceedings.

Farré-Maduell, E., G. González Gacio, S. Lima, A. Miranda-Escalada, and M. Krallinger. 2022. LivingNER Guidelines: Named entity recognition, normalization & classification of species, pathogens and food, April. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Grishman, R. and B. M. Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics.*

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

Guzman-Silverio, M., Á. Balderas-Paredes, and A. P. López-Monroy. 2020. Transformers and data augmentation for aggressiveness detection in mexican spanish. In *IberLEF@ SEPLN*, pages 293–302.

Kang, T., A. Perotte, Y. Tang, C. Ta, and C. Weng. 2021. Umls-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association*, 28(4):812–823.

Kim, J., Y. Kim, and S. Kang. 2022. Weakly labeled data augmentation for social media named entity recognition. *Expert Systems with Applications*, 209:118217.

Li, X., H. Zhang, and X.-H. Zhou. 2020. Chinese clinical named entity recognition with variant neural structures based on bert methods. *Journal of biomedical informatics*, 107:103422.

Liu, Q., P. Li, W. Lu, and Q. Cheng. 2020. Long-tail dataset entity recognition based on data augmentation. In *EEKE@ JCDL*, pages 79–80.

Luo, H. 2021. Emotion detection for spanish with data augmentation and transformer-based models. In *IberLEF@ SEPLN*, pages 35–42.

Luque, F. M. 2019. Atalaya at tass 2019: Data augmentation and robust embeddings for sentiment analysis. *arXiv preprint arXiv:1909.11241*.

Malmasi, S., A. Fang, B. Fetahu, S. Kar, and O. Rokhlenko. 2022. Semeval-2022 task 11: Multilingual complex named entity recognition (multiconer). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437.

Marivate, V. and T. Sefara. 2020. Improving short text classification through global augmentation methods. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 385–399. Springer.

Miranda-Escalada, A., E. Farré, and M. Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. *IberLEF@ SEPLN*, pages 303–323.

Miranda-Escalada, A., E. Farré-Maduell, S. Lima-López, L. Gascó, V. Briva-Iglesias, M. Agüero-Torales, and M. Krallinger. 2021. The profner shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health (SMM4H) Workshop and Shared Task*, pages 13–20.

Perera, N., M. Dehmer, and F. Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, page 673.

Raiman, J. and J. Miller. 2017. Globally normalized reader. *arXiv preprint arXiv:1709.02828*.

Rekabsaz, N., M. Lupu, and A. Hanbury. 2017. Exploration of a threshold for similarity based on uncertainty in word embedding. In *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings 39*, pages 396–409. Springer.

Rodriguez, S., R. Gretter, M. Matassoni, A. Alonso, O. Corcho, M. Rico, and F. Daniele. 2021. SmarTerp: A CAI system to support simultaneous interpreters in real-time. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 102–109, Held Online, July. INCOMA Ltd.

Schindler, D., F. Bensmann, S. Dietze, and F. Krüger. 2021. Somesci-a 5 star open data gold standard knowledge graph of software mentions in scientific articles. *arXiv preprint arXiv:2108.09070*.

Serrano, A. V., G. G. Subies, H. M. Zamorano, N. A. Garcia, D. Samy, D. B. Sanchez, A. M. Sandoval, M. G. Nieto, and A. B. Jimenez. 2022. Rigoberta: A state-of-the-art language model for spanish. *arXiv preprint arXiv:2205.10233*.

Tjong Kim Sang, E. F. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Wang, W. Y. and D. Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using % petpeeve tweets. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2557–2563.

Wei, J. and K. Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Wu, X., S. Lv, L. Zang, J. Han, and S. Hu. 2019. Conditional bert contextual augmentation. In *International conference* *on computational science*, pages 84–95. Springer.

Xie, Q., Z. Dai, E. Hovy, T. Luong, and Q. Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Yaseen, U. and S. Langer. 2021. Data augmentation for low-resource named entity recognition using backtranslation. *arXiv preprint arXiv:2108.11703.*

Zhang, X., J. Zhao, and Y. LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.