Named Entity Extractors for New Domains by Transfer Learning with Automatically Annotated Data

Emanuel Matos[1], Mário Rodrigues[2], and António Teixeira[1](✉)

[1] DETI/IEETA, University of Aveiro, Aveiro, Portugal
{easm,ajst}@ua.pt
[2] ESTGA/IEETA, University of Aveiro, Aveiro, Portugal
mjfr@ua.pt

Abstract. Named entity recognition (NER) tasks imply token-level labels. Annotating documents can be time-consuming, costly, and prone to human error. In many real-life scenarios, the lack of labeled data has become the biggest bottleneck preventing NER being effectively used in some domains and with some natural languages, with negative impacts on the quality of some tasks. To overcome the barrier of the lack of annotated data for new application domains in some natural languages, we propose a method that uses the output of an ensemble of NER's to automatically annotate the data needed to train a Bidirectional Encoder Representations from Transformers (BERT) based NER for Portuguese. The performance was assessed using MiniHAREM dataset with promising results. For domain relevant classes such as LOCAL, F1, Precision and Recall above 50% were obtained when training only with automatically annotated data.

Keywords: Named entities · Named Entity Recognition (NER) · Transfer learning · Automatic annotation

1 Introduction

Named Entity Recognition (NER) has application in several domains such as user interest modeling [27] and dialog systems [9], among other. Approaches to NER include training statistical sequential models based on handcrafted features, and more recently, deep learning models [8,14] to ease the burden of designing crafted rules. The models with robust performance are often created based on substantial amounts of labeled training data.

Because NER tasks imply token-level labels, annotating many documents can be time-consuming, and thus costly and prone to human error. In many real-life scenarios, the lack of labeled data has become the biggest bottleneck preventing NER being used in some domains and tasks. To solve the label scarcity problem, [10] proposes a method based on BERT [3] and distant supervision which involves

V. Pinheiro et al. (Eds.): PROPOR 2022, LNAI 13208, pp. 288–298, 2022.
https://doi.org/10.1007/978-3-030-98305-5_27

avoiding the traditional labeling procedure by matching tokens in the target corpus with concepts in knowledge bases such as Wikipedia and YAGO [15].

Our approach is also based on BERT. Instead of annotated data we use as reference the output of an ensemble of 3 general purpose NER developed previously [16] as well as an annotated dataset created by linguists named HAREM [17,20]. The performance was measured using a fully annotated corpus, MiniHAREM, that is not included in the first HAREM. The results show that the proposed method is feasible, and the resulting system was able to annotate previously unseen data.

This document is organized as follows: after this Introduction follows the second section elaborating on the relevant Related Work. In the third section the proposed Method is explained followed by Results in Sect. 4. The paper ends in Sect. 5 with the Conclusion.

2 Related Work

The set of approaches proposed for NER can be classified according to 2 types: Rule-based, covering both systems based in patterns and in lists (the so-called Gazetteers); and machine learning based.

Machine learning methods are more flexible to adapt to distinct contexts if there exists enough data about the target context. Diverse machine learning methods have been applied to NER, such as Support Vector Machines (SVM), Conditional Random Field (CRF) or Neural networks (NN) (e.g., [6]).

In recent years, Deep Learning and the existence of larger datasets resulted in relevant advances in NER with systems based on Long Short-Term Memory (LSTM), Bidirectional LSTMs (Bi-LSTM) and Transformers (e.g., [10,14]). LSTMs are recursive neural networks in which the hidden layers act as memory cells. As a result, they revealed better capabilities to deal with long range dependencies in data [7,8]. Transformers [23], introduced in 2017, are a deep learning model based on the attention mechanism designed to handle sequential input data, such as natural language. Their potential for parallelization enabled training using huge datasets. This created the conditions for the development of pretrained systems such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) [18]. Transformers demonstrated their superior efficiency in the recognition of named entities and in a variety of other classification tasks.

Following the general tendency, for Portuguese, the last years main developments on NER were systems featuring machine learning. A representative selection of works is presented in the following paragraphs:

The LeNER-Br system [13], presented in 2018, was developed for Brazilian legal documents. It features LSTM-CRF models trained using the Paramopama data set and achieved F1 scores of 97.0% and 88.8% for legislation and judicial entities, respectively. The results show the feasibility of using NER for judicial applications.

Pirovani and coworkers [19] adopted a hybrid technique combining Conditional Random Fields with a Local Grammar (CRF+LG) for a NER task at IberLEF 2019.

Lopes and coworkers [12] addressed NER for Clinical data in Portuguese showing Bi-LSTM and word embeddings superiority relative to CRF. They obtained F1 scores above 80%.

The first use of BERT in NER for Portuguese appeared in 2020 [21], adopting a BERT-CRF architecture combining transfer capabilities of BERT with the structured CRF forecasts. BERT was pre-trained using the brWac corpus (2.68 billion tokens), and the NER model was trained using the First HAREM and tested with MiniHAREM. This system surpassed the previous state art despite being trained with much less data.

To the best of our knowledge all these machine learning systems were trained with data annotated by humans, at least in a revision stage. This is a major limitation in the development of NER for new domains, which are in large demand by the expansion of potential application areas.

3 Method

The process used to assess the potential of automatically annotated entities is summarized in Fig. 1. Distinct BERT-based NERs were trained with automatically annotated data (BIO annotations) to detect words belonging to entities. The difference between both variants was if they included or not out of domain annotated entities, in this case obtained with the First HAREM dataset. The performance of entity detection of the 2 variants was assessed using a third annotated dataset (MiniHAREM) and by querying DBPedia. The main blocks of the processing are described in the following subsections.

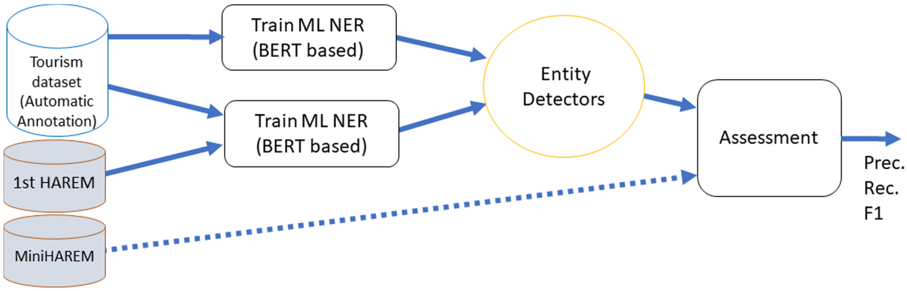
Tourism dataset (Automatic Annotation)
1st HAREM
MiniHAREM
Train ML NER (BERT based)
Train ML NER (BERT based)
Entity Detectors
Assessment
Prec.
Rec.
F1

Fig. 1. Overview of the process adopted to assess the potential of automatically annotated data for the development of Named Entity Detection for new domains. The development of Named Entity Detection for new domains implies the use of automatically annotated datasets for training machine learning models.

3.1 Datasets

Three datasets were used, two for training and one for testing, namely:

Automatically annotated dataset – we have used the results of our recent work on automatic annotation [16] as data to train the models. Briefly, a corpus for Tourism domain - a set of more than 300 documents obtained from Wikivoyage [25] - was annotated using a set of 3 NER (Linguakit NER [4,11], Alen NLP [1,5] and a DBPedia-based NER developed by the authors [16]). The output tags of these 3 NERs were combined to tag a word as part of an ENTITY or not, without including classification of the entity. More details can be found in [16].

HAREM datasets – Two HAREM [17,20] datasets were used, the First HAREM and the MiniHAREM, both with their labels annotated manually. The evaluation of the Entity detectors was performed against MiniHAREM. Profiting from the processing and XML made available by authors of [21][1], both text and BIO annotated files were derived.

The number of words annotated as part of Entity are presented in Fig. 2. The automatically annotated dataset for Tourism is the smallest one, being less than half the size of its combination with HAREM.

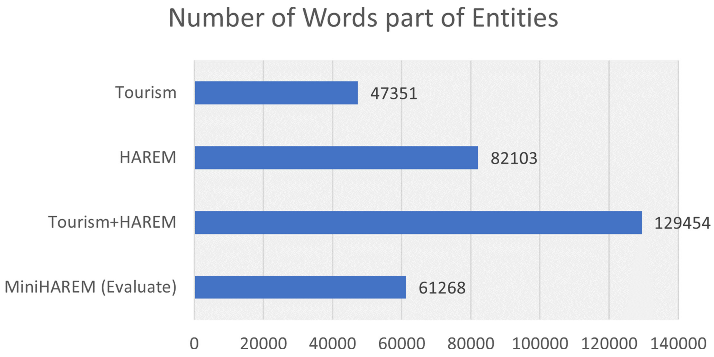

Fig. 2. Number of words annotated as part of an Entity for the 3 datasets used and the combination Tourism+HAREM used for training.

3.2 BERT-Based Classifiers for Entity Detection

Motivated by the recent evolutions highlighted in Sect. 2, a NER based on BERT [3] was selected for our experiments. After tests with several implementations, we have adopted a simple and documented implementation by Tobias Sterbak [22] using the Transformers package by Huggingface [26], Keras and TensorFlow. It uses a BERT case sensitive tokenizer – based on Wordpiece tokenizer – that splits tokens into subwords.

[1] https://github.com/neuralmind-ai/portuguese-bert/tree/master/ner_evaluation.

Two base systems were considered for our investigation: one trained using just automatic annotations (called Tourism); a second one trained with these automatic annotations plus the annotations of First HAREM (Tourism+HAREM).

Training

The models were finetuned with AdamW optimizer adopting the default parameters of [22], `lr` $= 3 \times 10^{-5}$ and `eps` $= 10^{-8}$, in a notebook with GPU NVIDIA GeForce RTX 2060.

The training datasets were split into training and validation, with 90% for training and 10% for validation. The stop criteria adopted were a maximum of 10 epochs or the increase of loss in the validation set.

The variation of loss with epochs is presented in Fig. 3 for the two variants trained. The training process stopped, for both variants, after the third epoch due to the loss increase. The number of epochs needed to finetune the model is aligned with the information in [22] that "a few epochs should be enough ... 3–4 epochs".

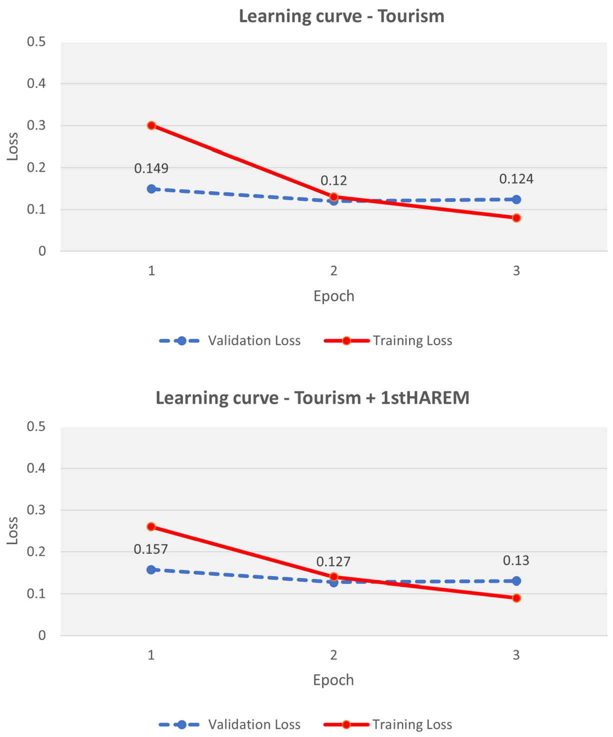

Fig. 3. Loss variation during the training of both system variants.

The information regarding processing time is presented in Fig. 4. It is visible that train and test durations are similar for both variants, and the complete process (train plus test) is under 8 min.

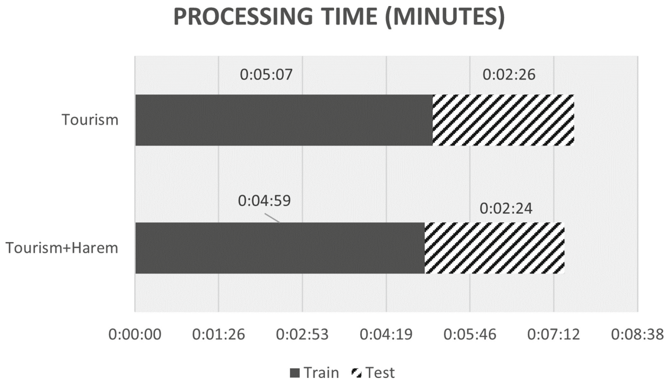

Fig. 4. Training and testing times (in minutes) for the two system variants.

4 Results

The output obtained by processing MiniHAREM is exemplified in Fig. 5. The performance was evaluated using the standard metrics (Precision, Recall and F1) considering both the individual words of the entity and the full entity as a single unit (the entity is well detected when all words composing it are correctly tagged).

The results obtained for both variants of the system (trained with and without using First HAREM) are summarized in Table 1.

Table 1. Results of the evaluation of both variants with MiniHAREM, showing word and entity metrics. Acc.* refers to Accuracy calculated excluding all words with tag “O”.

Train data	Word-based					Entity-based			
	Acc.	Prec.	Rec.	F1	Acc.*	Acc.	Prec.	Rec.	F1
Tourism only	92.6	90.0	37.3	52.7	35.8	88.6	41.4	26.4	32.2
Tourism + HAREM	94.0	94.0	49.6	64.9	48.0	88.6	56.3	40.4	47.1

Table 1 shows that even testing with an out-of-domain dataset:

- with automatically annotated, tourism domain, data the system could achieve a word-based precision and full entity detection of 90.0% and 41.4%, respectively.

Position	Word	HAREM Tag	BERT NER Output
4702	do	O	O
4703	Porto	B-LOCAL	B-ENT
4704	para	O	O
4705	Nine	B-LOCAL	B-ENT
4706	.	O	O
...			
4925	São	B-LOCAL	B-ENT
4926	João	I-LOCAL	B-ENT
4927	do	I-LOCAL	O
4928	Souto	I-LOCAL	B-ENT
4929	.	O	O

Fig. 5. Two fragments of the results obtained with the BERT-based NER using Mini-HAREM, showing the HAREM tag and the output of the variant trained only with automatic annotations.

– Results improved, as expected, by using first HAREM data to complement domain data in the training process.
– Recall is much lower. Over 50% of the words belonging to entities are not detected even when including HAREM data in the training. The best recall value when considering complete entities is 40.4%.
– The best full entity precision is approx. 56%.

As the test dataset contains several entity classes not annotated in our tourism data, it is worth looking in detail at the performance by class. The split of entity-based Precision, Recall and F1 by entity class is presented graphically in Fig. 6.

Bar-plots show that: 1) LOCAL presents the highest F1, being Precision and Recall above 50% when trained with just automatically annotated data. 2) PERSON obtains the second highest F1 but only when HAREM data is used to complement the domain data. 3) Classes that are not in the domain dataset such as WORK or ABSTRACTION present interesting Precision and Recall, showing some potential of the system to generalize.

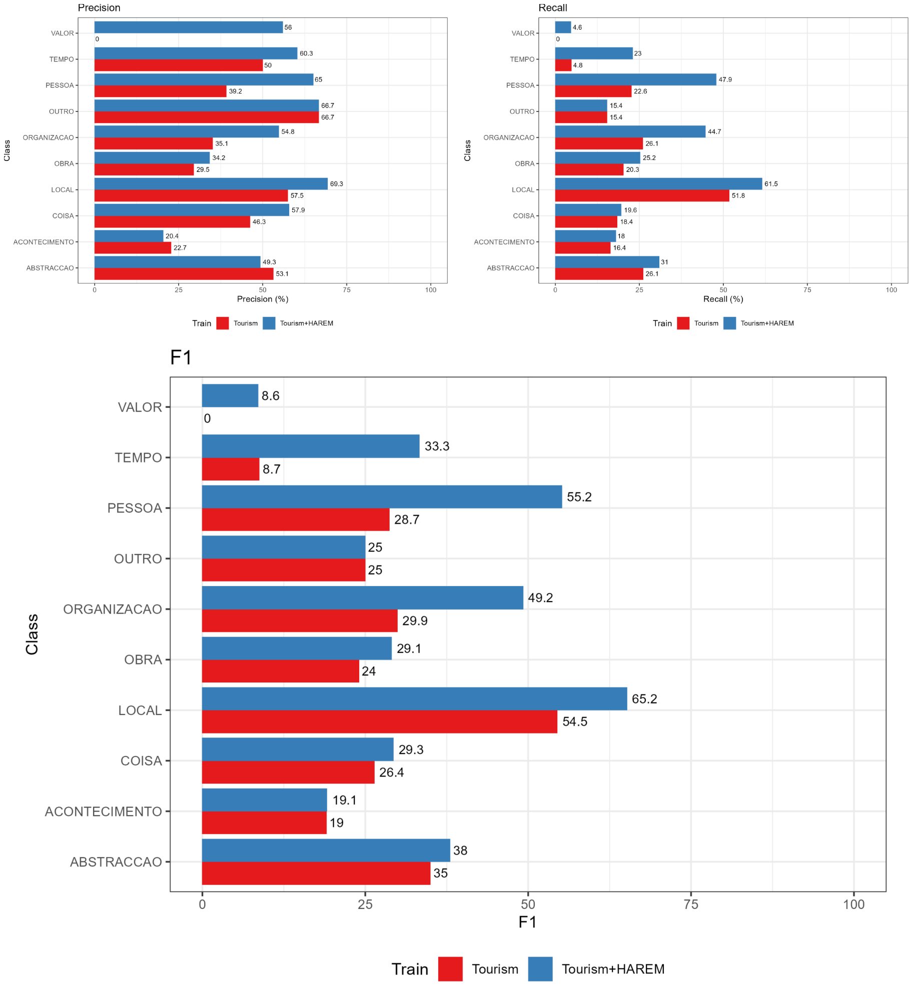

Fig. 6. Entity-based Precision (top left), Recall (top right) and F1 by entity class.

5 Conclusion

Aiming at making possible and simple creation of new NER for domains without annotated data available, **the potential of automatic annotation to provide the necessary datasets for training of entity detectors by fine-tuning of BERT models is assessed in this paper**.

A first proof-of-concept of the proposed method was evaluated with an existing manually annotated dataset, MiniHAREM. As the dataset used for testing - selected by being the only publicly available dataset for NER evaluation in Portuguese - is out-of-domain and integrates several entities without examples in the domain dataset, the results need to be considered with caution. Admitting limitations, the results are promising regarding the potential to create Named Entity detectors for a new domain without manually annotated data as demonstrated by the results obtained for classes more represented in the domain dataset (e.g., LOCAL). We also consider an interesting result the capability of the system to "generalize" from the training data and detect entities of classes that are not present in the domain data such as ABSTRACTION.

The main contribution of this paper is the proposal and evaluation of fine-tuning of pretrained BERT models with automatically annotated data to foster development of Named Entity Extraction for new domains.

5.1 Future Work

The interesting results obtained with a small domain dataset and (almost) off-the-shelf BERT-based model recommend the exploration of several improvements to the work presented, particularly:

- Extension of the Tourism dataset, by processing additional texts by the automatic annotation process (ensemble of NER).
- Addition of NER for classes such as TIME to the automatic annotation process.
- Exploration of recent evolutions in BERT-based NER and similar models (e.g., GPT-3 or FLAN) [2,24].
- Manual revision of part of the results obtained for a subset of Tourism texts to allow computing performance metrics for domain data.
- Try Transfer based Learning in the correction of decisions of the developed NER.
- Add the NEC post-processing step to classify the entity detected, using, for example, queries to DBPEdia and Wikipedia.
- Apply and evaluate the process proposed to new domains.
- Integrate the newly obtained NER in an Information Extraction pipeline.

Acknowledgement. This research was supported by IEETA Research Unit, funded by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UIDB/00127/2020.

References

1. Allen NLP - An Apache 2.0 NLP research library, built on PyTorch, for developing state-of-the-art deep learning models on a wide variety of linguistic tasks. https://github.com/allenai/allennlp
2. Brown, T.B., et al.: Language models are few-shot learners. arXiv:2005.14165 (2020)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
4. Gamallo, P., Garcia, M.: Linguakit: uma ferramenta multilingue para a análise linguística e a extração de informação. Linguamática **9**(1), 19–28 (2017)

5. Gardner, M., et al.: AllenNLP: a deep semantic natural language processing platform (2018)
6. Goyal, A., Gupta, V., Kumar, M.: Analysis of different supervised techniques for named entity recognition. In: International Conference on Advanced Informatics for Computing Research, pp. 184–195. Springer, New York (2019). https://doi.org/10.1007/978-981-15-0108-1_18
7. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
8. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arXiv:1508.01991 (2015)
9. Ketsmur, M., Teixeira, A., Almeida, N., Silva, S., Rodrigues, M.: Conversational assistant for an accessible smart home: proof-of-concept for portuguese. In: Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion, pp. 55–62. DSAI 2018, Association for Computing Machinery, New York, NY, USA (2018)
10. Liang, C., et al.: Bond: bert-assisted open-domain named entity recognition with distant supervision. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1054–1064. KDD 2020, Association for Computing Machinery, New York, NY, USA (2020)
11. Linguakit full (2020). https://linguakit.com/en/full-analysis
12. Lopes, F., Teixeira, C., Oliveira, H.G.: Contributions to clinical named entity recognition in portuguese. In: Proceedings 18th BioNLP Workshop and Shared Task (2019)
13. Luz de Araujo, P.H., de Campos, T.E., de Oliveira, R.R.R., Stauffer, M., Couto, S., Bermejo, P.: LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In: PROPOR. LNCS, Springer, New York (2018). https://doi.org/10.1007/978-3-319-99722-3_32
14. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1064–1074. Association for Computational Linguistics, Berlin, Germany (Aug 2016)
15. Mahdisoltani, F., Biega, J., Suchanek, F.M.: YAGO3: a Knowledge base from multilingual wikipedias. In: CIDR. Asilomar, United States (2013). https://hal-imt.archives-ouvertes.fr/hal-01699874
16. Matos, E., Rodrigues, M., Miguel, P., Teixeira, A.: Towards automatic creation of annotations to foster development of named entity recognizers. In: Queirós, R., Pinto, M., Simões, A., Portela, F., Pereira, M.J.A. (eds.) 10th Symposium on Languages, Applications and Technologies (SLATE 2021). Open Access Series in Informatics (OASIcs), vol. 94, pp. 11:1–11:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2021). https://drops.dagstuhl.de/opus/volltexte/2021/14428
17. Mota, C., Santos, D. (eds.): Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca (2008), http://www.linguateca.pt/LivroSegundoHAREM, iSBN: 978-989-20-1656-6
18. Patel, A., Arasanipalai, A.: Applied Natural Language Processing in the Enterprise: Teaching Machines to Read, Write, and Understand. O'Reilly Media, Incorporated (2021)
19. Pirovani, J.P., Alves, J., Spalenza, M., Silva, W., da Silveira Colombo, C., Oliveira, E.: Adapting NER (CRF+ LG) for many textual genres. In: IberLEF@ SEPLN, pp. 421–433 (2019)

20. Santos, D., Seco, N., Cardoso, N., Vilela, R.: HAREM: an advanced NER evaluation contest for Portuguese. In: Calzolari, N., et al. (eds.) Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC) (2006)
21. Souza, F., Nogueira, R., Lotufo, R.: Portuguese Named Entity Recognition using BERT-CRF (2020)
22. Sterbak, T.: Named entity recognition with BERT (2018). https://www.depends-on-the-definition.com/named-entity-recognition-with-bert/ last updated: 2020–04-24. Accessed on 24 Oct 2021
23. Vaswani, A., et al.: Attention is all you need. arXiv:1706.03762 (2017)
24. Wei, J., et al.: Finetuned language models are zero-shot learners. arXiv:2109.01652 (2021)
25. Wikivoyage. https://pt.wikivoyage.org/
26. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics (2020). https://www.aclweb.org/anthology/2020.emnlp-demos.6
27. Zhu, Z., Zhou, Y., Deng, X., Wang, X.: A graph-oriented model for hierarchical user interest in precision social marketing. Electron. Commer. Res. Appl. **35**, 100845 (2019). https://www.sciencedirect.com/science/article/pii/S1567422319300225