LANGUAGE-INDEPENDENT EMBEDDINGS FOR ENTITY RECOGNITION VIA LLM DATA-LEVEL KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Entity Recognition has always been one of the most important problems in Natural Language Processing. However, there wasn't much research aimed at creating a high-quality Multilingual Domain-Agnostic Foundation Model for Entity Recognition task. We introduce novel LLM-powered Data-Creation and Contrastive Learning-based Pre-Training procedures that enable us to create a new state-of-the-art Foundation Model for Entity Recognition task. It is designed and trained to have high performance on data coming from different domains and to enable language-independent features thanks to a new diverse multilingual training dataset. Our contribution surpasses all existing models on the English and Multilingual Entity Recognition tasks when used as a Foundation Model. We improved the macro F1-Score of multilingual BERT (Devlin et al., 2018) by 10 points on the single-language scenario and by 13.5 points on the multi-language scenario in French, German, English, Spanish, Italian, Polish, Portuguese, and Russian on the Entity Recognition Task. We open-source our model on the HuggingFace (Wolf et al., 2019) platform.

1 INTRODUCTION

Entity Recognition has always been one of the most important problems in Natural Language Processing. Since the raise of Deep Learning, the typical methodology of solving an Entity Recognition task consists of 2 steps:

1. Choosing a proper Foundation Model
2. Using computed features to make token/word-level predictions

Despite a lot of research papers published about Entity Recognition, most of the works attempted to get additional performance from step number 2 while neglecting the creation of a proper task-specific Foundation Model.

We argue that having a high-quality multilingual Foundation Model pre-trained for Entity Recognition on diverse data would enable researchers to get better results while requiring fewer labeled data points at the same time. Our contributions are as follows:

- Novel LLM-powered Data-Creation and Contrastive Learning-based pre-training procedures enabling efficient data-level LLM knowledge distillation and auto-labelling of any amount of data for model training
- Best open-source English Entity-Recognition Foundation Model
- Best open-source multilingual Entity-Recognition Foundation Model with language-independent features
- Optimal configuration for creating feature-based Entity Recognition models from open-source and our models

2 Related Work

There is a long history of existing foundation model that can be used to solve Entity Recognition Task. While RNN-based models (Sherstinsky, 2018) like ELMO (Peters et al., 2018) where popular during a specific time frame, we will be reviewing a more modern architectures that are currently used in the industry.

2.1 Pre-trained generic transformer-based models

Since the inception of pre-trained transformers (Vaswani et al., 2017), pre-trained BERT-like models became the most widely used foundation models. Such models as BERT(Devlin et al., 2018), RoBERTa (Liu et al., 2019) and others were created with the goal to be suitable for any kind of NLP task and their pre-trained objectives were good for general language modeling but heavily neglected task-specific traits and thus were not fully efficient for Entity Recognition, especially on the low-data scenarios. Moreover, in case of multilingual models, multilingual pre-trained models such as multilingual BERT and XLM RoBERTa (Conneau et al., 2019), according to our benchmark, didn't prove to be very efficient in Entity Recognition Task, especially in the case of the evaluation datasets where multiple languages were present at the same time both in train and in test split. We think that further adaptation of those models to Entity Recognition task is necessary.

2.2 Fine-tuned BERT-like models

In the recent years there have been several attempts at creating diverse Entity Recognition dataset that can be used to create your own model. For example, in NER-BERT (Liu et al., 2021), authors label an English Wikipedia subset with annotations covering over 300 entity classes. The problem with this data, however, is that the labeled texts come only from one source and thus the model will not be as good in more informal texts. When it comes to multilingual datasets, MultiNERD (Tedeschi & Navigli, 2022) is the most prominent example spanning over 10 languages and 15 entity classes. The problem, however, is that this data is not very informative because of only 1 text source and only 15 entity classes which does not allow for a great generalisation.

3 Data creation

While creating the dataset, it was important to create the one that spans over multiple languages, covers diverse sources and domain areas, and contains many entity classes.

In order to comply with source, domain and language diversity, we decided to use multilingual OSCAR dataset Ortiz Su'arez et al. (2020). This is the dataset of the raw filtered internet texts that were taken from diverse websites without limiting to only one source, texts are grouped by their language. Now we will present the intuition behind the auto-labelling procedure and the procedure itself.

Let's look at the example:

The World War II started on September 1 1939

As a human we can extract entities from this text without having a pre-defined set of classes:

- The World War II
  - Label: historical event
  - Description: a global war lasting from 1939 to 1945
- September 1 1939
  - Label: date
  - Description: date of the start of the war

Let's ask Large Language Model to do the same for any sentence with the following prompt:

> The goal is to create a dataset for entity recognition. Label as many entities, concepts, and ideas as possible in the input text. Invent new entity types that

> may not exist in traditional NER Tasks such as more abstract concepts and ideas. Make sure the entity concept is not part of speech but something more meaningful. Avoid finding meaningless entities.
>
> Output format (separate entities with new lines, everything, including description, and entity concept is written in **French**):
>
> entity from the text -—- entity concept -—- description of entity group/concept
>
> Input:

This prompt, while not looking very elegant, enables LLM to understand the task and extract high-quality diverse entities with correct and precise classes and descriptions. For each sentence we mention its language in the prompt in order to have language-consistency between the sentence and the description and concept language.

It's important to note that we used March 2023 version of GPT-3.5-turbo (Brown et al., 2020) instead of the June version as the June version often neglected the output format for some languages. Additionally, we want to highlight that specifying the desired output language is crucial because in 20% of German texts labeled and in 50%+ Spanish sentences labeled, format was not fully respected, requiring additional post-processing, or the output was in English.

In the OSCAR dataset, for each language, we remove all texts shorter than 100 and longer than 399 characters. We auto-label with the above-mentioned procedure 16 000 texts for each language (except 160 000 for English) out of English, French, German, Italian, Spanish, Italian, Polish, Portuguese, and Russian.

In the prompt we didn't limit the set of supported entity classes. It enabled us to have the most diverse and unique entity classes possible. Now, depending on the context, the same person can be an actor, a teacher or just a human. However, this also forced us to come up with a novel model training procedure as the conventional one (with a linear layer of embedding size by number of classes and a Cross Entropy loss) could not be applied here both due to the potentially unlimited number of classes and entity class identification inconsistency of LLMs.

4 Model Creation

4.1 Training procedure

In this subsecion we will explain the proposed training procedure on B input sentences (that can be represented as a $SentTokens$ matrix with size [B, $sentenceLen$] after tokenization) and $numDesc$ extracted pairs of label and description for those B sentences.

Let's concatenate entity concept and its description for each out of $numDesc$ entries. After concatenation, let's calculate the sentence embedding for each concatenated pair via the BERT-like model. The sentence embeddings are being calculated by the default methodology of the corresponding BERT-like model: CLS token, mean, min, etc. Thus we obtained a matrix $\boldsymbol{D}$ of size $numDesc$ by $embDim$.

We can obtain $\boldsymbol{L}$ (logit) matrix of size [B, $sentenceLen$, $numDesc$] by extracting token embeddings from the $SentTokens$ matrix and calculating a dot product between the embedded $SentTokens$ and $\boldsymbol{D}$. We use the same BERT-like model as of the sentence embedding before but he we use raw token embeddings to represent tokens.

Let's calculate a matrix $\boldsymbol{T}$ (target) of size [B, $sentenceLen$, $numDesc$] that is initialized with zeros. For each found entity, let's iterate over all B sentences, and for each occurrence of it, let's assign 1 in the $\boldsymbol{T}$ matrix for the tokens, present in the current entity.

Now, having $\boldsymbol{L}$ (logit) and $\boldsymbol{T}$ (target) matrices, we can train our model. In this paper, we explore several methodologies of transforming $\boldsymbol{L}$ into $\boldsymbol{T}$ approximation for loss calculation:

1. **BCE Loss & Sigmoid** where loss is computed only over the tokens that were found at least in one description according to $\boldsymbol{T}$ matrix. Before applying Sigmoid, we divide every value in the $\boldsymbol{L}$ matrix by a parameter that we call $temperature$.
2. **NLLLoss & Softmax over the $numDesc$ axis** and same loss compute constraint

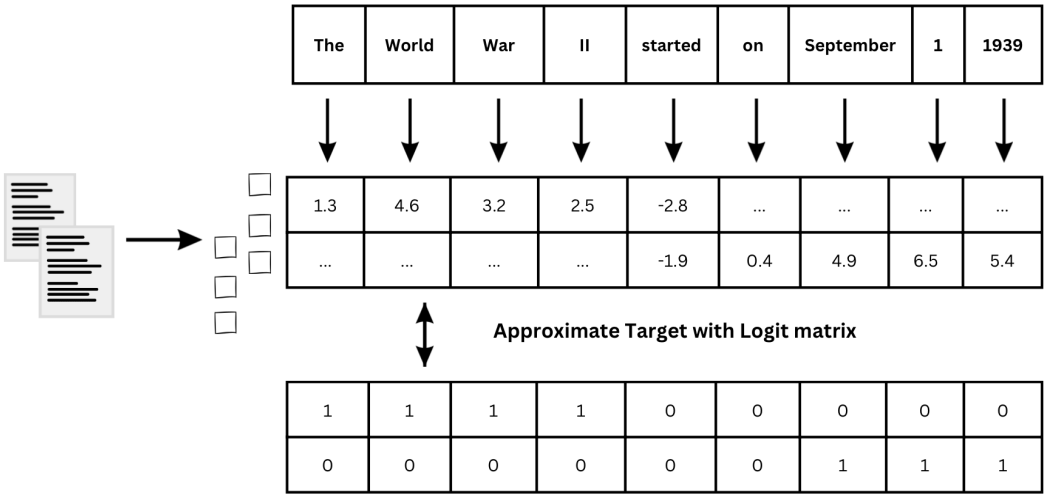

Figure 1: Training procedure example for batch size = 1 and the example from Data creation section

3. **(2) + multi-label NLLLoss & multi-label Softmax over $sentenceLen$ axis**

In all methodologies above, the loss is not computed over padding tokens.

General schema of the training procedure can be found on the figure 1.

To obtain the English model, we fine-tune the RoBERTa-base model over 10 epochs with the entire 160k English texts labeled, $lr = 0.0001$, Adam optimizer, linear scheduler with warm up of 1 epoch, and $temperature = 5$. We fine-tune only top 6 layers while keeping the bottom ones frozen. If not mentioned otherwise, we use BCE Loss & Sigmoid as the default training methodology. We choose the best checkpoint during training based on the performance on the validation split of the evaluation datasets, benchmarking is done at the end of each epoch. RoBERTa foundation model is used as it performs better than any other BERT-sized open-source model, according to our evaluations.

4.2 Multilingual Training Procedure Adjustments

When training a multilingual foundation model, we want a model that would output embeddings compatible between different languages and in this section we will talk about the methods that can be used to achieve it.

In this paper we compare two main training configurations:

- **only one language in a batch**: having only one language texts being present in each batch. This can be beneficial because we would introduce more difficult examples in a batch as, for example, it's pretty easy for the model to distinguish Russian description from English token
- **no constraints**: this mode is beneficial as model would learn how to model token embeddings jointly across all languages

We will also compare the following tricks to achieve embedding language-consistency:

- **Label and Description-based consistency**: having all descriptions and labels in the same language - English would potentially enable embedding language-consistency through contrastive learning process fueled by English label + descriptions embeddings only
- **Encoder-based consistency**: we propose to use additional encoder model to embed label + description pair. It potentially leads to embedding language-consistency through high-quality pre-trained language-independent sentence representations. As an encoder model, we will be using an E5 English/multilingual model (Wang et al., 2022).

- **English Teacher model-based consistency**: Finally, we also tried distilling the model with the teacher model obtained in the previous subsection from the English data. Following the work of Reimers & Gurevych (2020), we introduce 2 additional losses:
  - Model-level loss: mean-squared-error between embedding of the English label + description of the teacher and student models
  - Language-level loss: mean-squared-error between the embedding of the teacher model of the English label + description and the embedding of the student model for the original label + description.

  In both cases, we refer to the currently trained model as a student and to an English model from the previous subsection as a teacher. The role of the Model-level loss is to ensure the general high quality of the embeddings. The role of the Language-level loss is to make embeddings language-independent and suitable for multi-language scenarios. 0.2 was the best multiplier we were able to find in to use with those newly-introduced losses.

For **Label and Description-based consistency** and **English Teacher model-based consistency**, every label and description needed to also be translated into English, we do so separately and use corresponding model for each language pair of the source language and English by Helsinki-NLP research group (Tiedemann & Thottingal, 2020).

To obtain the multilingual model, we fine-tune the BERT multilingual base cased model (it outperformed other alternatives) over 15 epochs with the 16k text dataset for each language labeled, $lr = 0.00003$, and other parameters kept the same as for the English model.

5 RESULTS

5.1 EVALUATION PROTOCOL

Since our goal is to evaluate the foundation model's raw features,

1. Frozen model is used for the benchmark, no fine-tuning is allowed in order to test the inital foundation model's capabilities
2. Each evaluation dataset is split into train and test with 1000 examples in each split
3. Logistic Regression is trained over token embeddings, where targets are the real entity labels from the train split. We use IO Entity Recognition labeling methodology, no additional post-processing is applied to the model predictions.
4. F1-Score Macro is the main metric to benchmark model performance over all classes equally, without favouring the most popular ones

For evaluation in English, we use Conll 2003 (Tjong Kim Sang & De Meulder, 2003), MIT Movie (MIT, 2014a), MIT Restaurant (MIT, 2014b), Ontonotes 5 (Weischedel et al., 2013), and Bionlp 2004 Collier et al. (2004) datasets. Those datasets cover a quantity of domains and entity classes and enable fair evaluation. As a target English metric, we average the macro F1-Score over all English evaluation datasets.

For the multilingual model, we benchmark over French, German, Italian, Spanish, Portuguese, Polish, Dutch and Russian subsets of MultiNERD dataset.

We introduce two main multilingual testing scenarios:

- single-language scenario - Logistic Regression is trained and benchmarked only on one language at a time.
- multi-language scenario - Logistic Regression is trained and benchmarked on two languages at a time. In this benchmarking scenario, ideally, we would like to have the performance no less than an average of the single-language performances of the two taken languages. If the performance is noticeably lower, it means that our multilingual embeddings are not very compatible across languages

We noticed that adding output of the top frozen layer in addition to the last layer to Logistic Regression further boost the performance of the models, when use we refer to this trick as two emb.

Table 1: Comparison of training configurations

train method	target metric
(1): Sigmoid + two emb	0.7665
(2): Softmax + two emb	0.7662
(3): Double Softmax + two emb	0.7671

Table 2: Our English model vs existing models on Enlish benchmark

Model	**F1 Macro**
RoBERTa	0.7129
RoBERTa + two emb	0.7290
NER-BERT + two emb	0.7498
new NER-BERT + two emb	0.7563
ours (EN version) + two emb	0.7686

5.2 ENGLISH MODEL

5.2.1 LOSS FUNCTION COMPARISON

As mentioned before, there are three main training configurations we can apply:

- BCE Loss & Sigmoid
- NLLLoss & Softmax over the $numDesc$ axis
- (2) + multi-label NLLLoss & multi-label Softmax over $sentenceLen$ axis

According to our observations in the table 1, the difference between the training configuration is neglectable and thus we will be using a Sigmoid setup in the future.

5.2.2 OUR MODEL VS NER-BERT

Now, when the setup is clear, let's compare the proposed contrastive learning pipeline vs fine-tuning on the existing general-domain data.

To enable fair evaluation, we pre-trained our own NER-BERT models, enchancing original training methodology with our procedure of choosing the best model during training. We trained only top 6 layers of the model, in our case it led to better performance. Additionally, we auto-labeled raw NER-BERT's textual data with our auto-labeling procedure and trained a new model with our training methodology on it, we call it new NER-BERT.

As we can see in the table 2, original NER-BERT performs poorer than new NER-BERT with the only difference being the labeling process of data (while the textual data is the same): our auto-labeling procedure yields more informative labels than the human-crafted ones. Moreover, our model performs better than new NER-BERT, proving that labeling diverse raw internet data is more efficient that data coming from the cleaned version of Wikipedia. In this experiment we proved that our methodology is more capable than the most diverse and entity class-rich Entity Recognition dataset in the internet.

5.3 MULTILINGUAL MODEL

In this section, we will compare proposed multilingual methodologies on the single and multi-language evaluation scenarios. Initially, we will compare the "toy" models that were trained only on English, French, German and Spanish, and benchmarked on those languages and Dutch, and then we will present the overall performance of the final model. In the following experiments and bechmarks we use two emb trick during evaluation.

Firstly, let's compare training configurations and language-consistency tricks, we can see the results on the table 3.

Table 3: Comparison of different configurations and language-consistency tricks

Model	**Single-lang**	**Multi-lang**
mBERT	0.5483	0.4972
one lang in a batch (OLIB)	0.6076	0.5182
OLIB + en desc	0.6143	0.5163
OLIB + E5	0.61125	0.4871
OLIB + E5 fine-tune	0.6033	0.5125
OLIB + English KD	0.6098	0.5008
no constraints	0.603	0.568

Table 4: Singe and Multi-language performance of the final multilingual model

	de	es	fr	it	nl	pl	pt	ru
de	0.569	0.635	0.617	0.591	0.505	0.597	0.562	0.576
es	0.635	0.762	0.637	0.635	0.51	0.66	0.656	0.696
fr	0.617	0.637	0.593	0.593	0.582	0.632	0.548	0.601
it	0.591	0.635	0.593	0.614	0.505	0.651	0.602	0.603
nl	0.505	0.51	0.582	0.505	0.513	0.531	0.552	0.533
pl	0.597	0.66	0.632	0.651	0.531	0.616	0.546	0.609
pt	0.562	0.656	0.548	0.602	0.552	0.546	0.665	0.656
ru	0.576	0.696	0.601	0.603	0.533	0.609	0.656	0.653

In the table 3 OLIB stands for only one language in a batch configuration; en desc - for Label and Description-based consistency; E5 experiments - Encoder-based consistency; and English KD - for English Teacher model-based consistency methodologies.

We can see that all configurations outperform initial mBERT in the single-langauge scenario, and the majority of experiments outperforms mBERT in Multi-lang scenario. Moreover, all OLIB experiments perform much better in Single-lang scenarios but fall behind in a Multi-lang scenario. We also observe that all language-consistency tricks do not greatly contribute to the Multi-lang performance, however, they can be useful for increasing Single-language performance. In the end, given the small gap between Single and Multi-language performance, we decided to stick with no constraints run and train such a configuration on the full training dataset. The results are the following with more detailed metrics in the table 4.

English performance: 0.739

Multilingual Single-language performance: 0.6231

Multilingual Multi-language performance: 0.5936

In the table 4 on the diagonal we have the single-language performance for the corresponding language. Outside of the diagonal we have the multi-language performance of the model on the following pair of languages.

As we can see, the final model has great performance both in single-language and multi-language scenarios, and a great English performance.

When evaluated on the above-mentioned set of languages, the original mBERT's results are:

Multilingual Single-language performance: 0.5206

Multilingual Multi-language performance: 0.4578

As we can see, our model improved the macro F1-Score of multilingual BERT by 10 points on the single-language scenario and by 13.5 points on the multi-language scenario in French, German, English, Spanish, Italian, Polish, Portuguese, and Russian on the Entity Recognition Task. Such an improvement on the multi-language scenario proves that the language consistency was almost fully achieved.

6 CONCLUSION

The existing Entity Recognition foundation models either were not adopted to the Entity Recognition task directly or were limited by data sources/pre-defined entity classes in the training data. In this research we shared a procedure that enables the creation of the best BERT-sized multilingual domain-agnostic language-consistent foundation models for Entity Recognition task up to date.

We open source our models on HuggingFace.

P.S. The link to the HuggingFace model will be added after the anonymous review procedure to avoid potential deanonymization through the portal.

7 REPRODUCIBILITY STATEMENT

In this paper, we described the data processing both for training and evaluation datasets in full details. When using GPT-3.5-turbo's API, we made sure to stick to default parameters. During training of the foundation model, we fixed a random state of 42 for all frameworks used, enabled shuffle option in the train but disabled shuffle option in the test data loader. The same random state is also fixed during the evaluation of the final model. Evaluation data splits were pre-computed and saved in the hard drive to make sure they are the same for all experiments.

REFERENCES

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pp. 73–78, Geneva, Switzerland, August 28th and 29th 2004. COLING. URL https://aclanthology.org/W04-1213.

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. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Zihan Liu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. Ner-bert: a pre-trained model for low-resource entity tagging. *arXiv preprint arXiv:2112.00405*, 2021.

MIT. Mit movie dataset, 2014a. URL https://groups.csail.mit.edu/sls/downloads/.

MIT. Mit restaurant dataset, 2014b. URL https://groups.csail.mit.edu/sls/downloads/.

Pedro Javier Ortiz Su'arez, Laurent Romary, and Benoit Sagot. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pp. 1703–1714, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.156.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. URL http://arxiv.org/abs/1802.05365.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. *CoRR*, abs/2004.09813, 2020. URL https://arxiv.org/abs/2004.09813.

Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *CoRR*, abs/1808.03314, 2018. URL http://arxiv.org/abs/1808.03314.

Simone Tedeschi and Roberto Navigli. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 801–812, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.60. URL https://aclanthology.org/2022.findings-naacl.60.

Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003. URL https://www.aclweb.org/anthology/W03-0419.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2022.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. Ontonotes release 5.0. Text, Linguistic Data Consortium, Philadelphia, 2013. URL https://doi.org/10.35111/xmhb-2b84. Languages: English, Mandarin Chinese, Arabic, Chinese; Data Source(s): telephone conversations, newswire, newsgroups, broadcast news, broadcast conversation, weblogs, religious texts; Project(s): GALE; Application(s): information extraction, information retrieval; License(s): LDC User Agreement for Non-Members.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. URL http://arxiv.org/abs/1910.03771.