MEERQAT-IRIT at SemEval-2023 Task 2: Leveraging Contextualized Tag Descriptors for Multilingual Named Entity Recognition

Jesús Lovón-Melgarejo ♠, Jose G. Moreno ♠, Romaric Besançon ♢
Olivier Ferret ♢, Lynda Tamine ♠
♠Université Paul Sabatier, IRIT, UMR 5505 CNRS, Toulouse, France
♢Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{jesus.lovon, jose.moreno, tamine}@irit.fr
{romaric.besancon, olivier.ferret}@cea.fr

Abstract

This paper describes the system we submitted to the SemEval 2023 Task 2 Multilingual Complex Named Entity Recognition (MultiCoNER II) in four monolingual tracks (English, Spanish, French, and Portuguese). Considering the low context setting and the fine-grained taxonomy presented in this task, we propose a system that leverages the language model representations using hand-crafted tag descriptors. We explored how integrating the contextualized representations of tag descriptors with a language model can help improve the model performance for this task. We performed our evaluations on the development and test sets used in the task for the *Practice Phase* and the *Evaluation Phase* respectively.

1 Introduction

Named Entity Recognition (NER) (Grishman and Sundheim, 1996) consists in detecting groups of words as named entities from a given sentence and recognizing their type from an available list of entity tags. According to the entity tags list size, the NER task is classified as i) *coarse-grained* when the list is small (such as names of people, organizations, and locations, as proposed in the CoNLL task (Tjong Kim Sang and De Meulder, 2003)) or ii) *fine-grained* for a more extensive list (Ling and Weld, 2021). Additionally, we can distinguish different named entity types, referenced as traditional (person, location) and non-traditional (titles of creative work, such as a book or a song). Non-traditional entities, also called complex entities, are difficult to recognize due to problems such as semantic ambiguity (Ashwini and Choi, 2014).

The SemEval 2023 Task 2 Multilingual Complex Named Entity Recognition (MultiCoNER II) (Fetahu et al., 2023b) evaluates NER systems capable of identifying semantically complex and ambiguous named entities in 12 languages (English, Spanish, Hindi, Bangla, Chinese, Swedish, Farsi, French, Italian, Portuguese, Ukrainian, and German) providing multilingual resources to this end. Additionally, this task version proposes a fine-grained entity taxonomy with more than 30 different entity tags and a complex test set composed of low-context sentences divided into two subsets, named "noisy" and "clean". The "noisy" subset (available for eight languages) includes corrupted sentences that add noise on context tokens or entity tokens to simulate errors. MultiCoNER II proposes two types of evaluations: 12 monolingual tracks and one multilingual track. A task constraint is that multilingual models can not participate in the monolingual tracks, and similarly, the monolingual models can not participate in the multilingual track. Also, a specific dataset is provided for each phase of the evaluation: we adopt hereafter the term *development set* for the *Practice* phase and the term *test set* for the *Evaluation* phase.

Previous work identified complex entities in low-context sentences by focusing on augmenting the input query using external resources, training a knowledge base retriever (Wang et al., 2022), or using data-augmentation techniques (Gan et al., 2022). Most of these works concentrate on improving the contextualized representations of the query; however, less attention is given to the representations of the entity tags. We argue that contextualized representations of the entity tags can benefit the model's performance, particularly for a fine-grained entity taxonomy where some classes are usually underrepresented. Our intuition is to associate the input query entities with a clear definition of their entity tags, leading to an adapted representation space where entities sharing the same tag will have closer representations.

To this end, we defined rich and relevant contextualized *tag descriptors* to leverage the model representations for this task. We define a *tag descriptor* as combining a textual definition of the tag and a hypernymy-based definition to capture

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 878–884

the fine-grained taxonomic structure. We created hand-crafted definitions in multiple languages. Our system includes two different setups for the training and the test. For the training setup, we fine-tuned a Pre-trained Language Model (PLM), XLM-RoBERTa, computing contextualized representations for the query and the associated entity *tag descriptors* obtained from the golden annotations. Then, we aligned and aggregated both types of representations. The final token representations are then fed into a linear-chain Conditional Random Field (CRF) layer (Lafferty et al., 2001) for named entity prediction. For the test setup, we evaluate our model using only the improved PLM and the CRF layer. We tested our models in 4 monolingual tracks: English, Spanish, French, and Portuguese. Our experiments showed that injecting this knowledge helps enhance the model performance and shows consistent gain w.r.t the task baseline. Our systems obtained the rankings 21/34, 12/18, 13/17, and 14/17 for the English, Spanish, French, and Portuguese tracks, respectively.

2 Related Work

Broadly known PLMs, such as BERT (Devlin et al., 2018) and XLM-RoBERTa (Conneau et al., 2020), leveraged the performance on multiple NER benchmarks. This improvement has triggered the interest of the NLP research community by finding better setups to improve these models for the NER task. Notably, we can observe a significant presence of this type of model in the SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER) (Malmasi et al., 2022b). The MultiCoNER task provided a multi-lingual dataset (Malmasi et al., 2022a) focused on detecting semantically ambiguous and complex entities in short and low-context settings, which is more complicated than recognizing traditional entities (Meng et al., 2021). Based on this first version of the task, the MultiCoNER II task adds two more challenges by using a fine-grained entity taxonomy and simulating errors in the dataset.

Adding relevant context to the input phrase with external lexical resources, such as knowledge bases, helped to create improved token representations that positively impacted the performance of the MultiCoNER task (Wang et al., 2022). Similarly, previous works proposed pre-training enhanced-PLMs by fusing knowledge base representations with PLMs to improve token representations, which has been helpful in related NLP tasks (Zhang et al., 2019; Peters et al., 2019). However, the main drawback of these approaches is the high computing cost needed to pre-train the models, mainly due to the additional network architectures they contain. Another line of work explored the use of PLMs as knowledge bases to retrieve (Petroni et al., 2019) and inject facts (Talmor et al., 2020) of these external resources by translating them into textual utterances and applying to these utterances the pre-training task of the PLM. Based on these techniques, our approach considers a *tag descriptor* as the textual representation of a knowledge base fact. We aim to inject and retrieve this information helpfully into a PLM under a light training setup.

3 Data

The MultiCoNER 2 dataset (Fetahu et al., 2023a) comprises a fine-grained tagset of 33 tags grouped into six coarse-level categories (Location, Creative Work, Group, Person, Product, and Medical). The organizers provided 13 dedicated datasets for the 12 monolingual and multilingual tracks with train, valid, and test splits. Each dataset follows the CoNLL format, and the tag labels a BIO scheme.

We generated the *tag descriptors* using the following procedure. First, we identified each fine-grained tag with its corresponding coarse-level class in taxonomy. Then, we added definitions with two widely used taxonomies in the research community, WordNet (Miller, 1995) and ENE (Sekine et al., 2002), as references. If multiple definitions were available, we selected the one whose hypernym matched the coarse-level name class. We used as a complementary resource the Wikipedia category website when the category was found in none of the primary sources. We created the definitions initially in English. Then, we used an automatic translation tool to obtain their corresponding Spanish, French, and Portuguese definitions. Finally, we manually curated these translations to keep the orthographic correction and the contextual sense of the translation. Table 1 shows some examples of the generated *tag descriptors*.

4 Methodology

In this section, we describe our NER system, which comprises three modules: the Pre-trained Language Model, the Tag-Descriptor Encoder, and the Information Fusion modules, as shown in Figure 1. Our approach trains a PLM with the architecture of

Category	Lang.	Tag Descriptor
Medicine-Symptom	EN	A symptom is any sensation or change in bodily function that is experienced by a patient and is associated with a particular disease. A symptom is a medical term.
	ES	Un síntoma es cualquier sensación o cambio en la función corporal que experimenta un paciente y que se asocia a una enfermedad concreta. Un síntoma es un término médico.
	FR	Un symptôme est toute sensation ou modification d'une fonction corporelle ressentie par un patient et associée à une maladie particulière. Un symptôme est un terme médical.
	PT	Um sintoma é qualquer sensação ou mudança na função corporal que é experimentada por um paciente e está associada a uma determinada doença. Um sintoma é um termo médico.
Location-Human Settlement	EN	A human settlement is community of people smaller than a town. A human settlement is a location.
	ES	Un asentamiento humano es una comunidad de personas más pequeña que una ciudad. Un asentamiento humano es un lugar.
	FR	Un établissement humain est une communauté de personnes plus petite qu'une ville. Un établissement humain est un lieu.
	PT	Um assentamento humano é uma comunidade de pessoas menor do que uma cidade. Um assentamento humano é um lugar.

Table 1: Tag descriptors created for the fine-grained classes Symptom and Human Settlement in English (EN), Spanish (ES), French (FR), and Portuguese (PT).

Figure 1 to improve and adapt the model representations for the task but we rely only on the PLM to perform the evaluation. We take as input a sentence composed of n tokens $t = \{t_1, t_2, ..., t_n\}$ for the Pre-trained Language Model and Tag-Descriptor Encoder. The Tag-Descriptor Encoder first expands the input by concatenating the entity mention and the hand-crafted entity *tag descriptors*; then, it computes their contextualized representations. The Tag-Descriptor Encoder provides such a representation for each entity identified in the input sentence. We align these subtokens representations with the output representation of the Pre-Trained Language Model and compute an aggregated representation using the Information Fusion module. The combined representations are passed to a CRF layer that produces the label predictions.

Pre-Trained Language Model

Given an input sentence t, we use a PLM to compute the contextualized representation $\{t'_1, t'_2, ..., t'_n\}$ of tokens. Our system uses the XLM RoBERTa (XLM-R) model because of its good performance in this task. XLM-R is a multilingual version of the RoBERTa model pre-trained in over 100 languages. In particular, we use the available Hugging Face version (*xlm-roberta-large*).

Tag-Descriptor Encoder

From the annotated input of the training split, we identify the entity mentions and their annotated entity tag. We then expand each entity mention by concatenating the *tag descriptor* to it. The *tag descriptor* includes information based on the tag's definition and the fine-grained taxonomy's hypernymy. By adding the textual hypernymy definition, we also search to leverage the knowledge base encoder capability of these models. Finally, the entity tag name and the hand-crafted definition create an input with the form: $\{e_1, e_2, ..., w_1, ..., w_n\}$. We feed this input to another XLM-R to compute the tokens representations and we select only the tokens belonging to the entity mention $\{e'_1, e'_2, ...\}$.

As an example, Figure 1 considers the query q=*"1967: the man who fell to earth (producer with Michael Deeley)"*. This query has two recognized entity types: *Visual Work (Vis)*, which is a type of *Creative Work*, and *Artist (Art)*, which is a type of *Person*, corresponding to the subtexts *"the man who fell to earth"* and *"michael deeley"*, respec-

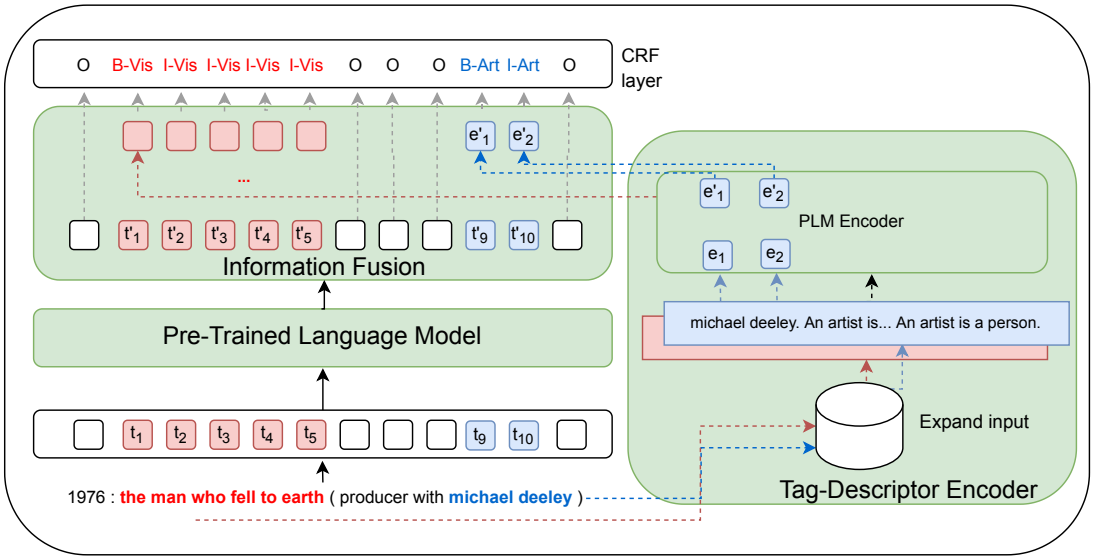

Figure 1: Illustration of the model architecture.

tively. The Tag-Descriptor Encoder generates two phrases by concatenating the *tag descriptors* to the entities. Specifically, the generated phrases are q1=*"the man who fell to earth. A visual work is [definition]. A visual Work is a creative work"*, and q2=*"michael deeley. An artist is [definition]. An artist is a person"*. Subsequently, we feed the PLM encoder with q1 and q2. Finally, we obtain the corresponding token representations for the subtexts.

This module aims to improve the token representations by integrating contextual information from the annotated tags. We do not train the elements in the Tag-Descriptor Encoder because the obtained representations were considered as informative.

Information Fusion

We compute an ensemble representation using the Pre-Trained Language Model and Tag-Descriptor Encoder modules. First, for each entity mention, we align the token computed from the Tag-Descriptor Encoder into the output representations of the PLM. We added zero-vectors for subtoken that do not belong to entities. We then perform an average weighted sum to compute the final representations:

$$(1 - b) * t'_i + b * e'_i \quad (1)$$

where b is a hyperparameter with values between $[0, 1]$.

Finally, we fed the CRF layer with the Information Fusion representations to predict the best entity tag sequences from all possible sequences.

5 Experiment and Results

Experimental Setup

We trained four models, one for each monolingual track: English, Spanish, Portuguese, and French. We used values of $b = 0.15$ for training, with a learning rate of $2e-5$ for Spanish, French, and Portuguese, and $1e-5$ for English, with a batch size of 32. We trained for 5 epochs and used an Nvidia RTX6000 graphic card. Our training time was around 3 hours per model. We used a value of $b = 0$ for testing.

5.1 Results

We analyzed our results for both evaluation sets, *development* and *test*. We submitted four monolingual models in their corresponding tracks during the evaluation. At the time of the competition, there were no official baseline scores. We implemented non-official baselines with the base model and hyperparameters described by the task organizers.

For the *development* test set, we considered two baselines. A first non-official baseline was obtained by an XLM-RoBERTa model with a CRF layer[1] and default settings for its hyperparameters (noted *XLM-R**) and another one by an optimized model (noted *XLM-R*) with the best hyperparameters for the respective languages. We used the same number of epochs for all models for fair comparisons. However, using *tag descriptors* showed more gap in performance at the first training steps.

[1]The scores are obtained from https://github.com/modelscope/AdaSeq/tree/master/examples

Model	EN	ES	FR	PT
XLM-R*	60.7	65.0	61.4	63.9
XLM-R	62.3	66.1	64.7	65.6
Our model	62.2	67.1	65.0	65.9
Δ / XLM-R*	+1.5	+2.1	+3.6	+2.0
Δ / XLM-R	-0.1	+1.0	+0.3	+0.3

Table 2: Macro-averaged F1 scores obtained for the development set in the four tracks EN, ES, FR, and PT.

Table 2 shows the performance of these models for the *development* test set and the improvements made by our model. We first compare our model with our general baseline *XLM-R** through the Δ / *XLM-R** row. Our model shows improvements in all languages, with a more significant impact on the French and Spanish tasks, up to 3.6 points on the F1 score, and a lower impact on the English task. Similarly, we compare our model with a more robust baseline *XLM-R*. Except for the English track, we still observe improvements in all languages up to 1.0 points on the F1 score. Furthermore, in Table 3, we identified the fine-grained classes that show constant improvement for all the languages. Our results showed an improvement of up to 28.6 points on the F1 macro score obtained in different languages. These results show that our model can incorporate the provided tag information in all the tested languages. However, a significant drawback of our approach is that we obtained zero scores for fine-grained tags unseen in training, showing an important limitation in relying on annotations from the training set.

For the *test* set, we include the best model in the competition for each track, named the *Top-1* model, considered an upper bound. Scores computed by the task organizers in the Evaluation Phase for the test set are reported in Table 4. We also added

Class	EN	ES	FR	PT
OtherLOC	+3.6	+1.7	+8.6	+3.6
HumanSettl	+2.35	+2.2	+0.7	+2.3
Station	+7.0	+10.6	+8.8	+7.0
MusicalWork	+3.1	+0.5	+3.9	+3.1
WrittenWork	+0.1	+5.3	+1.5	+0.1
OtherPER	+3.9	+2.9	+3.6	+3.9
Symptom	+28.6	+1.7	+5.8	+28.6

Table 3: Fine-grained tags improved by our model in comparison with our baseline XLM-Rob.

	Our model	Top-1 Model
Track EN-English		
clean	60.5	88.1
noisy	54.7	79.8
overall	58.7	85.5
Track ES-Spanish		
clean	64.9	91.7
noisy	58.8	85.8
overall	62.9	89.8
Track FR-French		
clean	61.3	91.6
noisy	53.7	85.1
overall	58.9	89.6
Track PT-Portuguese		
clean	61.9	87.3
noisy	56.2	83.4
overall	60.0	86.0

Table 4: Macro-averaged F1 scores obtained in the four mono-lingual tracks for the three types of the test set (clean, noisy, and overall).

the scores reported for the two subsets, noisy and clean. The final ranking is based on the overall macro F1. The scores obtained for the *test* set are similar to our scores for the *development* test, showing that our models did not overfit. Although our results suggest that injecting relevant context information of the entity tag into the models helps to improve performance, more is needed to outperform other models that use external resources to leverage the model representations. Additionally, our model is hindered by input simulation errors, decreasing its performance by 6.9%, 7%, 8.8%, and 6.3% on each of the English, Spanish, French, and Portuguese tracks, respectively. However, this impact is globally fairly the same for our model as for the *Top-1* model.

6 Conclusion

This paper describes our approach to improving token representation for the NER task. We evaluated the effectiveness of injecting context-relevant information about tag descriptors to improve the token representations of a PLM. The results for the development and test sets showed important improvements over the baseline when considering tag context information. Considering that the

Top-1 model requires extensive computation and textual resources for training and testing, we consider that our proposition, that only requires very limited additional knowledge, can be adopted as a first step towards developing better approaches for low-resource languages. In future work, we will evaluate how to extract better representations from the hierarchical fine-grained taxonomy and adapt this improved version of XLM-RoBERTa as a base model to other approaches, such as the best-performing method of this evaluation.

Acknowledgments

This work has been supported by the MEERQAT project (ANR-19-CE23-0028), granted by ANR the French Agence Nationale de la Recherche. This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011012638R1 made by GENCI.

References

Sandeep Ashwini and Jinho D Choi. 2014. Targetable named entity recognition in social media. *arXiv preprint arXiv:1408.0782*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023a. MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023b. SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Weichao Gan, Yuanping Lin, Guangbo Yu, Guimin Chen, and Qian Ye. 2022. Qtrade AI at SemEval-2022 task 11: An unified framework for multilingual NER task. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1654–1664, Seattle, United States. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Xiao Ling and Daniel Weld. 2021. Fine-grained entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1):94–100.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. 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, Seattle, United States. Association for Computational Linguistics.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512, Online. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

F. Petroni, T. Rocktaschel, A. H. Miller, P. Lewis, A. Bakhtin, Y. Wu, and S. Riedel. 2019. Language models as knowledge bases? In *In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*, 33:20227–20237.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. 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*, pages 142–147.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, Wei Lu, and Yong Jiang. 2022. DAMO-NLP at SemEval-2022 task 11: A knowledge-based system for multilingual named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1457–1468, Seattle, United States. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.