Challenges in including extra-linguistic context in pre-trained language models

Ionut-Teodor Sorodoc[*] **Laura Aina**[*] **Gemma Boleda**[*] [†]

[*]Universitat Pompeu Fabra
[†]ICREA
Barcelona, Spain
`{firstname.lastname}@upf.edu`

Abstract

To successfully account for language, computational models need to take into account both the linguistic context (the content of the utterances) and the extra-linguistic context (for instance, the participants in a dialogue). We focus on a referential task that asks models to link entity mentions in a TV show to the corresponding characters, and design an architecture that attempts to account for both kinds of context. In particular, our architecture combines a previously proposed specialized module (an "entity library") for character representation with transfer learning from a pre-trained language model. We find that, although the model does improve linguistic contextualization, it fails to successfully integrate extra-linguistic information about the participants in the dialogue. Our work shows that it is very challenging to incorporate extra-linguistic information into pre-trained language models.

1 Introduction

Identifying the real-world entity an expression refers to is crucial for Natural Language Processing, since humans use language to talk about the world. This, however, requires models that represent the real world such that linguistic expressions can be mapped to them. For instance, in Figure 1, which is a snippet of a dialogue from the TV show *Friends*, we need to know that it is Joey Tribbiani who is speaking to be able to interpret the pronoun "I". State-of-the-art NLP models typically focus on linguistic context, not on extra-linguistic context such as who is speaking to whom. We aim at integrating extra-linguistic context, in particular information about participants in a dialogue; also, we aim at combining it with information coming from the linguistic context.

We focus on the character identification task of SemEval 2018 (Choi and Chen, 2018), aimed at classifying mentions from the dialogue scripts of the TV show *Friends* (see Figure 1). The model that won the SemEval competition (Aina et al., 2018) proposed an external module to encode entity information in a structured way (henceforth, "entity library"). This approach enabled the incorporation of extra-linguistic information, in particular speaker information, which allowed the model to learn patterns such as "*I* refers to the character that is speaking"; and, as a result, it worked comparatively well on rare entities. However, Aina et al. (2019) showed that the model's good performance was not correlated with meaningful entity representations. Moreover, the model performed poorly in expressions that require a good grasp of the linguistic context, like 3rd person pronouns and common nouns.

JOEY TRIBBIANI (183):
"*. . . see Ross, because I think you love her .*"
335 183 335 306

Figure 1: Example of the dataset. It shows the speaker (first line) of the utterance (second line) and the ids of the entities to which the target mentions (underlined) refer (last line).

Aina et al.'s base model was an LSTM trained from scratch on the character identification task (with the exception of pre-trained non-contextualized word embeddings). We propose to instead add the entity library to a pre-trained language model: BERT (Devlin et al., 2019). Pre-trained language models (Peters et al., 2018; Devlin et al., 2019) have been shown to provide good contextual representations (Bai et al., 2021), and they have enabled advances also in referential tasks (Joshi et al., 2020; Zhou and Choi, 2018; Yang and Choi, 2019). We expected that combining BERT with the entity library would synthesize the benefits of both, encoding and exploiting both the extra-linguistic and linguistic information in the context. We also expected that, as a result of these improvements, this model would yield better entity representations.

Contrary to expectation, however, we do not

Proceedings of the Third Workshop on Insights from Negative Results in NLP, pages 134 - 138
May 26, 2022 ©2022 Association for Computational Linguistics

improve on the state-of-the-art model of Aina et al. (2019). Through analysis, we show that our model does improve the performance for context-dependent expressions, such as third-person pronouns, suggesting that it is better at handling the linguistic context; however, it performs worse on expressions that depend on the extra-linguistic context, such as first- and second-person pronouns, which are much more frequent in the data. Moreover, the entity representations are only marginally improved. The problem, we argue, comes from the fact that integrating extra-linguistic information in pre-trained language models is far from trivial.

2 Method and main results

Task In order to have a comparable setup to previous studies, the dataset and the task are the same as the ones described in Choi and Chen (2018). The training and test data span the first two seasons of the sitcom *Friends*, and the task is to predict which character is referred to by each referring expression (see Figure 1).

Model In our model, the input tokens go through a pre-trained BERT. Then the speaker information (i.e., an embedding identifying the character who produced the utterance) is concatenated to the token representation. This representation is fed to a multi-layer perceptron (MLP). The output of this step is compared to the entity library (EntLib) proposed in Aina et al. (2018), via dot products with each character embedding in the EntLib, in order to produce the final prediction (softmax over the dot products). The entity library is a learnable matrix where each row is associated with one of the 401 characters from the dataset. As in the version in Aina et al. (2019), the parameters of the speaker embedding matrix and of the entity library are shared. The weights of BERT are tuned to the character identification task. Section A.2 in the Appendix reports model details.

The most notable differences of our architecture with that of Aina et al. (2018) and Aina et al. (2019) are the following: 1) We run the input text through a pre-trained language model; 2) our model processes the input token with its textual context before accessing the speaker information. By contrast, Aina et al.'s architecture directly passes the input token to the LSTM jointly with the speaker. This latter difference will be crucial in explaining the results, as we will see in the next section.

		all (78)		main (7)	
	models	F_1	Acc	F_1	Acc
random	-EntLib	40.4	63.6	70.6	69.4
	+EntLib	43.8	64.4	71.2	70.4
BERT	frozen-EntLib	31.6	64	72.5	72.8
	frozen+EntLib	35.3	63.8	70.9	71.1
	finet.-EntLib	38.6	62.2	68.9	69.1
	finet.+EntLib	**51.4**	*70.5*	*76.9*	*77.6*
LSTMEnt	+EntLib	49.6	**77.6**	**84.9**	**84.2**

Table 1: Model parameters and results on the character identification task. *finet*: fine-tuned.

We conduct ablation experiments to investigate the benefits of different components of our model:

- **random embeddings**: the BERT component is substituted by randomly initialized embeddings. Each token is linearly mapped to a vector, with no representation of sequences.
- **frozen BERT**: the BERT component of the model is not fine-tuned on the character identification task, and only the other components are updated during training.
- **-EntLib**: the model does not include the entity library. The output of the MLP is directly mapped to 401 dimensions to predict an entity.

Results The main results are presented in Table 1.[1] The newly proposed model does not improve over the best performing model from Aina et al. (2019): it is better on F1 score for all entities, and worse for the other three metrics. However, while Aina et al.'s model (henceforth, LSTMEnt) has the best overall results, it outperforms the proposed model (fine-tuned BERT +EntLib, henceforth BERTEnt) only on a few kinds of expressions, as shown in the analyses in Section 3.

Table 1 also shows that the entity library improves over all 3 model variations, confirming that dedicating a specialized component to entity representation is helpful for referential tasks. Among our variants, the complete model (BERTEnt) is the best, showing that all the components are beneficial for the task. The models initialized with random embeddings are comparable to the models with frozen BERT embeddings. This suggests that BERT representations are not directly applicable to the current task, without being adjusted through fine-tuning; that may be due to the differences between the data

[1] While the prediction is over 401 entities, "all entities" in Table 1 are only 78 because this is the number of entities appearing in the test data.

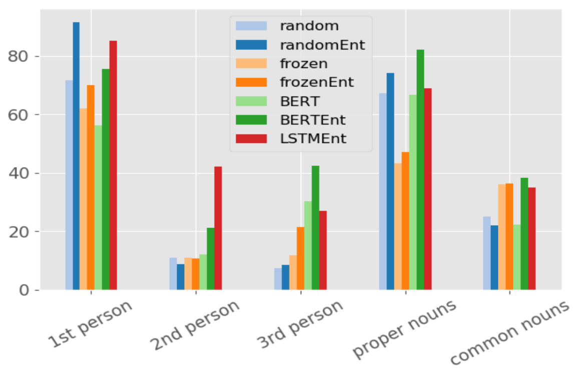

Figure 2: F1-score by type of referring expression (setup: all entities).

BERT was trained on (mostly narrative text) and the data we are deploying it on (dialogues from TV sitcoms).

3 Why does BertEnt not improve results?

Figure 2 presents the F1-score for the analyzed models for different types of referring expressions: first/second/third-person pronouns, proper nouns and common nouns. The graph shows results corresponding to all entities (column 'all' in Table 1). A graph focusing on the main entities is included in the Appendix.

As for first-person pronouns, recall that their interpretation depends on extra-linguistic information (who is speaking). Our models have speaker embeddings; to learn the right generalization, they should map the "I" token to the relevant speaker embedding. The entity library facilitates this process, and, accordingly, it is a beneficial component for first-person pronouns across all models.

Moreover, this is a type of referring expression that is easy for the models. The best strategy is actually to learn to treat the token representation for a first-person pronoun as a constant that functions simply as a prompt for the speaker embedding. This explains why the best results are actually obtained with random embeddings and entity library: The other models (including LSTMEnt) contextualize tokens, changing them depending on the content of the message. Since first-person pronouns do not depend on the linguistic context, but only on the extra-linguistic context, the other models have a harder time learning the right mechanism.

Second- and third-person pronouns are remarkably difficult for all models, and we find contrasting results between BERTEnt and LSTMEnt. BERTEnt is much worse than LSTMEnt at second-person pronouns, which again need extra-linguistic information (who the addresse is). As we explain in more detail later, in this case the problem is that in the current architecture speaker information is not contextualized together with the linguistic context. Instead, BERTEnt is better than LSTMEnt for third-person pronouns. This behaviour is expected given that third-person pronouns are tokens that require contextualization in the linguistic context (not the dialogue participants), and BERT specializes in contextualized representations.

Proper nouns are rigid designators, such that no contextual information is needed to predict which character "Ross" refers to (at least in the context of the sitcom) – neither linguistic nor extra-linguistic information. What is needed is to map the proper nouns to the corresponding characters, something that again is facilitated by the entity library. Most models are able to learn this mapping, with the exception of models with frozen BERT, which cannot adapt their proper noun representations to the context of the sitcom. BERTEnt is instead the most successful model for proper nouns, surpassing even LSTMEnt.

And the performance of BERTEnt is similar to that of LSTMEnt. This result is unexpected because common nouns bear resemblances to third-person pronouns (requiring contextualization, e.g. in the case of "woman") and to proper nouns (with some being more associated to a given character, like "paleontologist" with Ross), and BERTEnt outperforms LSTMEnt in both. However, common nouns are difficult for all the models. This can be traced back to two factors: 1) common nouns are rare in the training data; 2) the models are not learning good entity representations, which is necessary to learn the associations between nouns and characters (such as "paleontologist" with Ross). See Appendix A.5 for model biases that depend on training data distribution, and A.6 for the quality of entity representations.

Overall, the results show that BERTEnt and LSTMEnt have complementary strenghts: BERTEnt is better at accounting for linguistic context (with best results in third-person pronouns and proper nouns), and LSTMEnt at extra-linguistic context (with best results in first- an second-person pronouns). However, LSTMEnt achieves the best overall accuracy (Table 1) because of the data distribution: 44.4% of the datapoints are first-person pronouns, and 27.9% are second-person pronouns.

Thus, our proposed model succeeded in achieving better linguistic contextualization, but failed in incorporating extra-linguistic information, in particular information about the participants in the dialogue. We believe that the issue is that pre-trained language models like BERT do not have a "space" for extra-linguistic information; thus it is difficult to add it to current architectures. In particular, recall that, in our model, the speaker embedding is added at the output level: each token is processed by BERT, and then the speaker embedding is concatenated to the token. This means that the speaker embedding is not contextualized in the linguistic input, except via the MLP that further maps the concatenated token+speaker embeddings to the final decision. In LSTMEnt, instead, the token and the speaker embedding are processed jointly by the language model.

To understand the implications of this, consider the case of second-person pronouns: the entity we refer to when we use "you" is most probably an interlocutor who is the speaker of previous or future utterances. The current architecture doesn't have a straightforward way to access this information.

The way to go would be to include speaker information directly in the architecture of BERT. Since this entails all kinds of technical and conceptual issues, and in the spirit of "recycling" language models for referential tasks, we tried a middle-ground solution. We added a self-attention layer on top of the concatenation of the token and speaker information.[2] The self-attention layer operates on the whole sequence given as input: it compares the hidden representation at time step t with the hidden representations at all the other time steps. These comparisons are used to create a weighted representation. This layer should lead to incorporation of interlocutor information into the current representation. It however didn't work as expected: in our hyperparameter search, the best models did not use this component. This could be due to the component lacking a recency feature that encourages the model to focus more on the speakers surrounding the current token. For instance, for expressions like "you", the referent is usually a participant in the vicinity of the current utterance, such that it is harmful to consider all the spans considered in the BERT processing layer (more than 100 in the best instantiations of the model). Even though positional embeddings offer the possibility of focusing on more recent tokens, this information might not reach the output of BERT; thus the issue here could again be the fact that we include speaker information after BERT processing.

4 Conclusion

Our initial hypothesis was that the proposed model, BERTEnt, would attain the same performance as the previous state-of-the-art model (LSTMEnt) on mentions requiring extra-linguistic information, while improving linguistic contextualization and possibly the encoding of entity information. We instead find that the model does improve in linguistic contextualization (cf. higher performance in third-person pronouns), but instead fails to integrate extra-linguistic information about the participants in the dialogue (cf. lower performance in first- and second-person pronouns). Also, BERTEnt only slightly improves over LSTMEnt on entity representations (see Appendix A.6). The entity library does continue to be a valuable module, as in previous work (Aina et al., 2018, 2019), boosting performance across the board. Future work can focus on studying the benefits of the entity library in other pretrained models.

These results highlight requirements for successful architectures in situated Natural Language Processing. A model should be able to dynamically switch, depending on the input, between a strong sensitivity to the linguistic context and to the extra-linguistic context, to capture, e.g., that "I" points to the speaker, while "she" is to be disambiguated using the discourse context. This requires models to integrate the extra-linguistic context in their representations, a capacity that is severely underdeveloped at the moment. We have tackled the specific case of the participants in a dialogue, and have shown that it is very challenging to incorporate this kind of information in pre-trained language models. In order to address this issue, a possible approach for future research would be to develop a model which extends BERT to a multi-modal two-stream model, specialized on dialogue.

The *Friends* data that we have used is small for deep learning standards; one obvious way to go is to use more task-specific training data. Also, future work needs to conduct experiments on other dialogue-oriented tasks, in order to confirm our conclusion.

However, training data on any given "world", such as that of a particular TV show, or the envi-

[2] We tried 1/2/4 attention heads and 1/2 layers of attention.

ronment in which an artificial assistant is typically deployed (think Siri or Alexa), is inherently limited, such that newer models will need to be able to do more with less.

Acknowledgments

We are grateful to the anonymous reviewers for their valuable comments. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154). We are also grateful to the NVIDIA Corporation for the donation of GPUs used for this research. This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains.

References

Laura Aina, Carina Silberer, Ionut-Teodor Sorodoc, Matthijs Westera, and Gemma Boleda. 2018. AMORE-UPF at SemEval-2018 task 4: BiLSTM with entity library. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 65–69, New Orleans, Louisiana. Association for Computational Linguistics.

Laura Aina, Carina Silberer, Ionut-Teodor Sorodoc, Matthijs Westera, and Gemma Boleda. 2019. What do entity-centric models learn? insights from entity linking in multi-party dialogue. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3772–3783, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiaxin Bai, Hongming Zhang, Yangqiu Song, and Kun Xu. 2021. Joint coreference resolution and character linking for multiparty conversation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 539–548, Online. Association for Computational Linguistics.

Jinho D. Choi and Henry Y. Chen. 2018. SemEval 2018 task 4: Character identification on multiparty dialogues. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 57–64, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Zhengzhe Yang and Jinho D. Choi. 2019. FriendsQA: Open-domain question answering on TV show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197, Stockholm, Sweden. Association for Computational Linguistics.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.