Probing the representations of named entities in Transformer-based Language Models

Stefan F. Schouten and **Peter Bloem** and **Piek Vossen**
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
{s.f.schouten,p.bloem,p.t.j.m.vossen}@vu.nl

Abstract

In this work we analyze the named entity representations learned by Transformer-based language models. We investigate the role entities play in two tasks: a language modeling task, and a sequence classification task. For this purpose we collect a novel news topic classification dataset with 12 topics called RefNews-12. We perform two complementary methods of analysis. First, we use diagnostic models allowing us to quantify to what degree entity information is present in the hidden representations. Second, we perform entity mention substitution to measure how substitute-entities with different properties impact model performance. By controlling for model uncertainty we are able to show that entities are identified, and depending on the task, play a measurable role in the model's predictions. Additionally, we show that the entities' types alone are not enough to account for this. Finally, we find that the the frequency with which entities occur are important for the masked language modeling task, and that the entities' distributions over topics are important for topic classification.

1 Introduction

The probability a language model should assign to a sequence depends not only on what is being said, but also on the context, i.e. who is saying it, where, when, and why? Some types of context such as (cultural) background knowledge may already be represented to some degree within pre-trained language models. However, recent work shows that when it comes to world knowledge language models and knowledge bases are complementary, and that various forms of integration are beneficial (Safavi and Koutra, 2021). Being able to condition on the context explicitly would be particularly useful when we consider, for example, more specific cultural knowledge or interpersonal knowledge, which are unlikely to be contained in pre-training corpora.

In order to integrate language models and knowledge bases effectively it is important to know precisely how these two sources of information complement each other. Entities are specifically interesting as they occur within characteristic contexts learnable by language models and at the same time provide access to knowledge graphs.

In this work we investigate what information is present in the entity representations of Transformer-based (Vaswani et al., 2017) language models. We also study whether some entities' representations contain more information than others and why. Finally, we show how much these aspects change after a pre-trained language model is fine-tuned.

For our investigation we choose News articles as our primary source of data. News articles often describe events that involve numerous entities. Which is the primary reason they have been used in the past for entity-related tasks such as NERC (Tjong Kim Sang and De Meulder, 2003), NEL (Hoffart et al., 2011), and coreference resolution (Pradhan et al., 2012). We fine-tune and evaluate our models for news topic classification and masked language modeling. For this purpose, we collect a new dataset of English news articles by following links cited on Wikipedia pages covering many newsworthy incidents (Vossen et al., 2018). Our data collection method allows for this dataset to easily be expanded with additional topics and languages in the future. Furthermore, the set of entities linked to from Wikipedia pages covering such incidents can also provide us with a 'shortlist' of entities likely to be referenced in the news articles themselves. These entity links will allow integration with Wikidata, an avenue we wish to explore in future work.

Our analysis includes two complementary methods: diagnostic models (Veldhoen et al., 2016; Adi et al., 2016; Conneau et al., 2018) and a novel method of analysis we call entity mention substitution. *Diagnostic models* are trained on a relevant task (in our case entity recognition) with hidden representations of another model as input. The di-

Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 384 - 393

agnostic model is kept as simple as possible such that its performance can be attributed to the information being available in the hidden representation. *Entity mention substitution* measures the impact of various kinds of substitutions on the prediction of the model. If the impact is high we interpret this as evidence that the entity was important for the prediction. By manipulating which entities we choose as substitutes and by comparing the results to those of the diagnostic models, we answer the following research questions:

RQ1: When entities are mentioned in the input text, are they identified and used by Transformer-based language models?

RQ2: Does a Transformer-based language model either partially, or fully represent entities by their type?

RQ3: Do the answers to RQ1 & RQ2 depend on: **(a)** the frequency with which an entity-mention occurs in the data; and **(b)** the distribution of that entity across the news topics?

We make two important contributions. First, we collect a novel news classification dataset we call RefNews-12, which consists of 106,167 articles which cover 9,878 incidents grouped by 12 topics. Second, we analyze what information is present in entity representations in two ways. One concerns training and evaluating diagnostic models on entity recognition using only the model's hidden representations for entities as input. The other involves corrupting entity mentions in the data in various ways, showing how the model relies on the entities to make predictions.

We find that entities are identified in pre-trained models both before and after fine-tuning. Entities are also used to perform the task for which the model is trained or fine-tuned, even if they cannot be identified clearly by their representation. We also find that entities are represented by more than their type. Finally, our experiments suggest that the importance of the frequency with which entities occur and entities' distribution over topics is task-specific.

2 Related Work

The use of news articles to study the interaction of entities and topics is not new. Newman et al. (2006) mention that "news articles are ideal because they have the primary purpose of conveying information about who, what, when and where." We use them for the same reason, but focus on studying the entity representations in recent Language Models.

Previous investigations into the representation of entities in Language Models have come from various directions.

Broscheit (2019) investigates entity knowledge in pre-trained BERT through entity linking. They frame entity linking as a token classification problem over the entire vocabulary of 700K entities, thereby solving mention detection, candidate generation, and entity disambiguation simultaneously. When trained with BERT's weights frozen this method still obtains decent F1 scores (67.8 versus SotA of 85.8 by (Zhang et al., 2021)) on the AIDA benchmark (Hoffart et al., 2011). This indicates that BERT already assigns representations that are sufficiently distinct for entity linking to a lot of the entity tokens. We expect that this distinctness does not necessarily imply that entities are really treated by the model as individuals. Thus, we directly investigate the degree to which these entity representations are interchangeable.

Sorodoc et al. (2020) study whether pre-trained language models capture information helpful with the resolution of pronominal anaphora. They hypothesize that the model will learn helpful grammatical properties, but not semantic-referential information. To test this hypothesis they train diagnostic models and analyze how their performance varies as the variables of interest change. Their evidence suggests that language models do in fact learn some referential aspects, but that they are still much better at grammar. We also investigate the presence of semantic properties in representations of entities, but do so with different methods and include models that have been fine-tuned.

Biswas et al. (2021) use entity embeddings obtained from various language models to classify entities as one of 14 types. Interestingly, BERT embeddings obtain the lowest accuracy of the models tested. Where Biswas et al. (2021) embeds only the name of the entity, our work studies representations of entities that appear in context.

A number of other works do not investigate the representations of entities specifically, but test to what extent Language Models are able to reproduce relational world knowledge, which involves numerous facts about entities as well (Petroni et al., 2019; Roberts et al., 2020). For a recent survey of

this type of research see Safavi and Koutra (2021).

3 Methodology

The following section describes the method by which we collect our data (3.1), and the two methods we used to perform our analysis (3.2, 3.3).

3.1 Dataset Collection

We collect a novel news topic classification dataset based on articles that are linked to from Wikipedia. For our investigation we prefer data that is categorized across a large number of (hierarchical) topics, which can be used to construct datasets of varying difficulties.

For this purpose we use the Multilingual Wiki Extraction Pipeline (Vossen et al., 2020). This tool takes as input a set of Wikidata Event Types and queries Wikidata for each type's set of incidents. For example, for the Wikidata Item 'homicide' (Q149086) the pipeline finds items that are instances of homicides, i.e. all items that link to it with the 'instanceOf' property. The incidents that we select are those Wikidata Items for which a time and place are known. These incidents' Wikipedia pages are then scraped for links to news articles. Besides the articles linked on the Wikipedia pages, we also include the page itself. The links to other Wikipedia pages can be used to supervise Named Entity Recognition and Entity Linking.

To obtain the initial set of Wikidata Event Types, we make use of IPTC's Media Topics standard. This standard consists of a hierarchical taxonomy of terms intended for use by media to categorize their productions. Along with the hierarchy of topics, IPTC also distributes a mapping of these topics to Wikidata. We query the Wikidata Event Types referenced in the mapping to obtain the number of incidents available for each topic.

The dataset we collect for the experiments in this paper are based on a selection of 12 diverse topics, each with a number of incidents that is manageable but sufficient. We call this dataset RefNews-12. See Table 1 for an overview of the selected topics, their Wikidata ID, the number of incidents, and the total number of articles we scraped.

RefNews-12 is based on news articles from a wide variety of publications, none of which we obtained (or attempted to obtain) permission from to redistribute their work. To circumvent this legal obstacle, we do not directly distribute the articles themselves, but rather a set of URLs. To further increase the reproducibility, each URL is also accompanied by the timestamp of a 'capture' in the Internet Archive's Wayback Machine from which we obtained our copy. This set can be used by any interested party to obtain a dataset near-identical to that used for the experimentation in this work. This set of URLs for the articles which constitute RefNews-12 can be found at `https://github.com/sfschouten/refnews`, along with instructions and code to collect the dataset.

3.2 Diagnostic Models

Diagnostic models[1] (Veldhoen et al., 2016; Adi et al., 2016; Conneau et al., 2018) are used to investigate if the representations learned by a system include information about some feature of interest. The diagnostic model is trained to predict this feature from the representations. Its architecture is chosen to be as simple as possible, which allows for the diagnostic model's performance to be attributed to the information in the representation.

3.3 Entity Mention Substitution

Substituting the entities mentioned in the data allows us to establish whether entities are important for news topic classification and masked language modeling. It also tells us whether entity representations capture the the entity's type. Finally, we use it to investigate if either of these things depend on the frequency of the entity in the data, or the entity's distribution over the classes.

The core of this method involves measuring the effect of the substitutions on the final prediction. If a model's prediction consistently does not change after substitution then clearly the original entities' representations are not meaningfully different from the substitute representations. By identifying a few key ways in which entities can be (dis)similar, and substituting such that one particular property is either changed or kept the same, we can test if that property is present in the model's representations.

Specifically, we hypothesize that the effect of a mention's replacement depends on at least the following variables.

Type Equality Whether or not the original and substitute entities are of the same type. If type is present in the representations, then substituting by entities of the same type should give better performance than if we substitute for random entities.

[1] Also known by various other names, including 'diagnostic classifiers', 'auxiliary prediction tasks' and 'probing tasks'.

IPTC Name	Wikidata ID	#Incidents	#Articles
homicide	Q149086	938	25,123
natural disaster	Q8065	1,103	10,900
referenda	Q43109	722	5,627
transportation accident and incident	Q11822042	1,822	16,551
sport event	Q167170	518	7,188
coup d'etat	Q45382	339	3,416
educational testing and examinations	Q27318	1,352	8,168
record and achievement	Q1241356	2,352	15,034
armed conflict	Q350604	137	4,155
sports transaction	Q18515440	196	4,018
primary election	Q669262	193	2.612
transport	Q7590	206	3,375
Total		9,878	106,167

Table 1: RefNews-12: topics, number of incidents and articles.

Frequency How frequently the original and substitute entities occur in the training data. We expect that the embeddings associated with more frequently occurring entity mentions will have acquired more distinctive representations during training, and thus have a greater impact on the model's predictions.

Topic Shift How much difference there is between the distribution over topics of the original and substitute entity mentions. For example, if 'entity1' is only mentioned in articles of topics A and B and we replace it with 'entity2' which is only mentioned in articles of topics C and D; then we would expect that to have a greater impact than if we had replaced 'entity1' with an entity that is mentioned in a collection of articles with similar topics.

4 Experiments

This section details our experimental setup.

4.1 The (fine-tuned) language models

We use DistilBERT as our model of choice for all experiments. This decision was made because of resource constraints, specifically because we train multiple instances for each setting in order to calculate model uncertainty. DistilBERT is a 40% smaller distilled version of BERT (Devlin et al., 2019). While much smaller, it retains much of BERT's original performance. Choosing this model allows for a smaller computational budget.

We fine-tune DistilBERT seven times on RefNews-12 for both news topic classification and masked language modeling. Each time the classification head's parameters are initialized using a different seed. For topic classification we also train a model with the same architecture but from initialization (rather than using pre-trained weights).

All models are trained with a batch size of 72. The base learning rate is set to 0.0005, and subject to 2000 steps of warmup followed by a linear decay. They are evaluated on the validation split every 500 batches and training is stopped early if the performance does not improve 5 times in a row.

4.2 NER diagnostic models

We train diagnostic models on the CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003). We choose this dataset, because like RefNews-12, it consists of news articles. This procedure reveals to what degree the hidden representations can be used to predict which tokens are part of named entities. We train them for a pre-trained DistilBERT and the seven instances of DistilBERT fine-tuned on RefNews-12 for both tasks. A different classifier is trained for each layer of each model instance, revealing at what layer entities are most clearly represented. To put the results of the diagnostic classifiers in perspective, we also train them on an untrained randomly initialized model. The diagnostic classifiers are trained with the same hyperparameters as above, but without a learning rate warmup.

4.3 Replacing entity mentions

We perform a series of experiments where we replace entity mentions that occur in RefNews-12's news articles. We do not have gold-standard entity mention labels for all of RefNews-12 (only the Wikipedia pages have mention annotations through hyperlinks), so we use an additional DistilBERT model[2] that has been fine-tuned for Named Entity Recognition and Classification to obtain silver-standard labels instead.

Replacing the entity mentions may result in two kinds of changes in the model's predictions. First, our intervention may cause the model to confidently predict something else, this might mean that a different entity consistently causes a different topic to be predicted. However, if the manipulated inputs are sufficiently different from the training distribution, they may also cause greater model uncertainty, making predictions more arbitrary. We can use our independently seeded instances to differentiate between these two scenarios. The seven independently seeded model instances can be thought of as samples from an approximate posterior over the model's weights (Gustafsson et al., 2020). Thus, we use the variation in the predictions of these seven instances to approximately measure model uncertainty. Specifically, we evaluate the uncertainty using the method suggested in (Lakshminarayanan et al., 2017), which is to sum the KL Divergence between each model instance's prediction and the average of those predictions.

4.3.1 [MASK] Token Baseline

In this first baseline we replace entity mentions by the [MASK] token. This prevents the model from being able to use the information captured by the entity representations directly. However, both BERT and DistilBERT's training objective included predicting masked-out tokens in English text. Therefore the model may be able to reconstruct some of the missing information. Thus, we expect this to have relatively little effect on the performance and uncertainty of the pre-trained and fine-tuned models.

4.3.2 Random Token Baseline

The second baseline involves the mentions being replaced by random tokens. In this case the model has to identify the tokens that are out of place first, before it has the option of ignoring them. Thus we expect a somewhat larger effect on model performance. Contrary to the first baseline we expect this second baseline to come with significant model uncertainty, because this intervention should produce inputs the model did not see during training.

4.3.3 Random Mention

In this variant we substitute entity-mentions by a different randomly selected mentioned entity. With this substitution, the model may have a harder time identifying and ignoring the substitution, because other entities will not seem particularly out of place compared to random tokens. Therefore, if the named entities are important to complete the task at hand, and their representations are meaningfully different, we would expect the model to confidently predict something else. This would look like a large shift in the prediction where the shift is similar for each model instance (low uncertainty). If the model's performance is comparable to the baselines, we interpret this as evidence that either all entities are represented more or less the same way, or their differences are ignored in practice.

4.3.4 Type Invariant

The next step is to replace mentions only by others of the same type. If even entities of the same type are still represented in meaningfully different ways, we expect the performance to stay below the baselines. This would be evidence that entities are represented distinctly even within their type. However, if performance is comparable to the baselines, we interpret this as evidence that entities must be represented no more distinctly than their type.

4.3.5 Most Frequent

The final substitution we make is based on the frequency with which entity mentions occur in the data. We select the substitute mentions from the 100 most frequently occurring entities. If substituting for more frequently occurring entities affects the performance more than substituting for random entities, then the most frequent entities' representations must have encoded more information relevant to the task.

4.3.6 Correlation with shift in frequency and topic distribution

Finally, we calculate the correlation between two metrics and the loss of each model. For the first metric we calculate the difference in log-frequency

[2] https://huggingface.co/elastic/distilbert-base-cased-finetuned-conll03-english

between the original and substitute entities, averaged over each substitution per sequence. For the second metric we calculate how each entity is distributed across the topics. We then calculate the KL divergence of the original distribution from the substitute distribution, also averaged over each substitutions per sequence.

5 Results

Using the experimental results we can now answer our research questions. Figure 1 shows the results of the entity mention substitution experiments. Figure Figure 2 shows the accuracies obtained by the diagnostic classifiers. Table 2 shows the correlation coefficients obtained described in 4.3.6.

In Figure 1a we can see a significant drop in mean accuracy between both *original* and random-tokens (from 84.1% / 85.9% to 74.7% / 78.4% for From init. / Fine-tuned respectively, both with $p<0.001$), and between *random-tokens* and *random-mention* (from 74.7% / 78.37% to 70.4% / 75.99% with $p<0.001$ / $p=0.008$). Replacing a mention with another mention leaves a sentence that is more coherent than when it is replaced with random tokens. Despite this, we observe lower accuracy for *random-mention*. It seems that for topic classification the model is capable of ignoring random tokens, but cannot do the same for the random mentions. Instead, the model's predictions are considerably different with the substitute entity mentions, decreasing the accuracy as a result. From the model uncertainty in Figure 1b we can see that the drop in accuracy is not caused by increased uncertainty (uncertainty decreases from 0.150 for *random-tokens* to 0.107 for *random-mention*). We interpret this as evidence that the model uses entity mentions in its prediction.

Unfortunately, we cannot conclude the same from the masked language modeling results in Figure 1c. For this task the performance does not worsen going from *random-tokens* to *random-mention* (from 5.80 / 4.37 to 4.95 / 2.82). We also cannot make the same argument when comparing between *mask* and *random-mention*, because although the performance does deteriorate (from 3.73 / 1.83 to 4.95 / 2.82), this may also be explained by the uncertainty going up (from 0.098 to 0.147, no uncertainty for pre-trained). However, results from the diagnostic classifiers (Figure 2) do indicate that the identification of entities is beneficial for masked language modeling, since their performance increases compared to the Random and Pre-trained baselines.

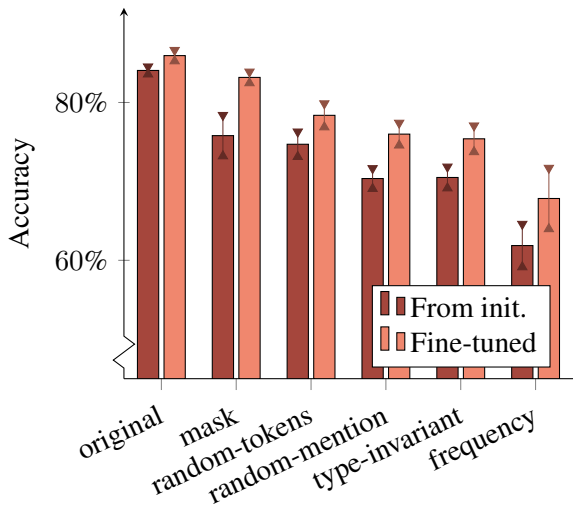
Furthermore, the diagnostic classifiers indicate that entities are identified in pre-trained and fine-tuned language models to a much greater degree than in models trained from initialization for topic classification.

In conclusion, entities are identified and used by the fine-tuned models for the topic classification task. However, for models trained from initialization entities are not easily identifiable from their representations. Despite that, their presence is still used by the model to perform the topic classification task. For masked language modeling we only have evidence of them being identified, but not of them being used. Thus, the answer to **RQ1** (*"When entities are mentioned in the input text, are they identified and used by Transformer-based language models?"*) is that entities are identified by language models, but whether they are used in practice depends on the task that the model is fine-tuned for.

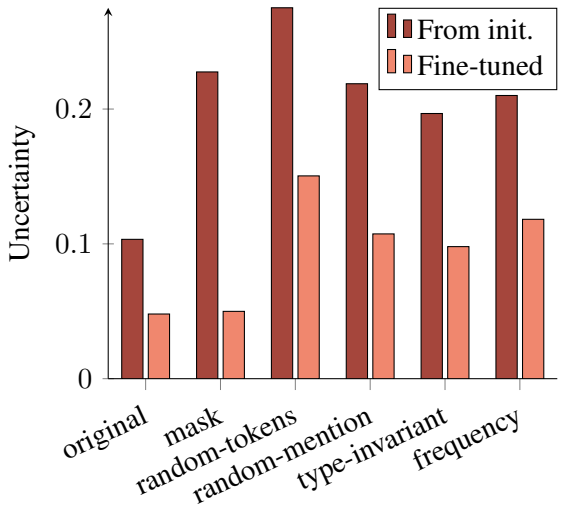
Looking at the *type-invariant* substitution in Figure 1a we can see that there is no significant difference in accuracy compared to the *random-mention* substitution. By comparing to *random-tokens* however, we can see the same pattern as we saw for *random-mention*: accuracy and uncertainty are both down (accuracy from 74.7% / 78.4% to 70.5% / 75.4%, uncertainty from 0.275 / 0.150 to 0.197 / 0.098). So even when substituting for mentions of the same type the model is still confidently changing its prediction, indicating that type is at least not the only aspect being looked at when predicting topics.

In setting out to answer **RQ2** (*"Does a Transformer-based language model either partially, or fully represent entities by their type?"*) we have not been able to present new evidence indicating that type is used by Transformer-based language models, but we have demonstrated that entities are not generally represented only by their type.

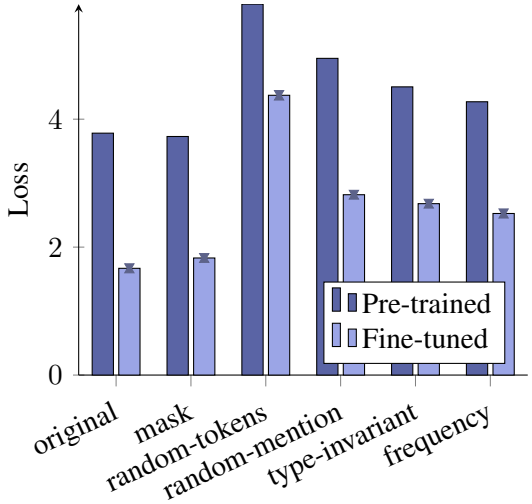
The substitution by the most frequently mentioned entities for the topic classification task as seen in Figure 1, shows a drop in performance compared to *random-mention*, but this is paired with a (modest) increase in uncertainty. Thus, the results of this particular experimental setting are inconclusive. However, in Table 2 we can see that there are dependencies between the performances



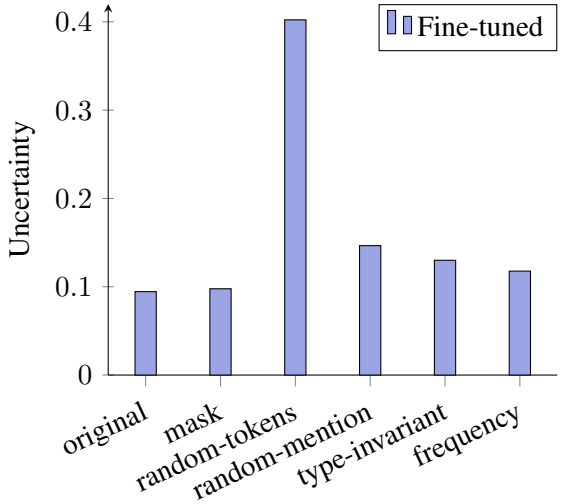
(a) Accuracy for News Topic Classification.



(b) Uncertainty for News Topic Classification.



(c) Loss for Masked Language Modeling.

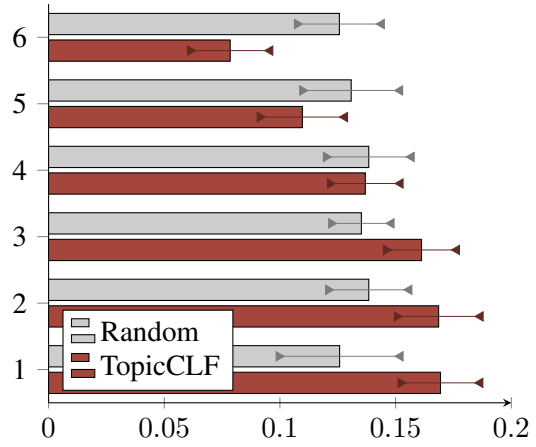


(d) Uncertainty for Masked Language Modeling.

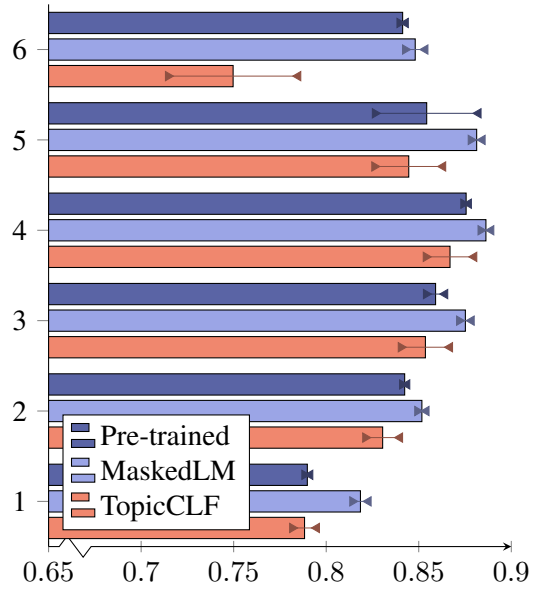
Figure 1: Performance metrics and uncertainty estimates obtained while performing Topic Classification and Masked Language Modeling for our entity-mention substitution experiments using our RefNews dataset. Error Bars display 95% confidence intervals indicating sensitivity to random initialization.

Task	Training	Variable	random-mention	type-invariant
TopicCLF	From init.	Frequency	0.00 ± 0.01	-0.01 ± 0.02
		Topic	0.10 ± 0.01	0.17 ± 0.01
	Fine-tuned	Frequency	0.00 ± 0.01	-0.02 ± 0.01
		Topic	0.10 ± 0.01	0.19 ± 0.01
MaskedLM	Pre-trained	Frequency	0.15	0.19
		Topic	-0.03	0.00
	Fine-tuned	Frequency	0.07 ± 0.00	0.13 ± 0.00
		Topic	-0.03 ± 0.01	-0.02 ± 0.01

Table 2: Pearson correlation between difference in frequency/topic and the model's loss while performing masked language modeling or topic classification for our entity-mention substitution experiments.



(a) Randomly initialized (untrained) model compared to topic classifier trained from initialization.



(b) Pre-trained and fine-tuned models.

Figure 2: Diagnostic classifier F1 score (x-axis) on NER for each layer (y-axis) of various models. Error bars display 95% confidence intervals indicating sensitivity to random initialization (of the diagnostic model, and in the case of the fine-tuned models also the model being probed).

obtained on either task and the average difference in frequency and topic distribution. For the models trained on MaskedLM when the difference between the frequency of the original and substitute entities increases so does the loss of the model. The same is true for the difference in topic distribution on the model fine-tuned for topic classification. Therefore our answer for **RQ3** (*"Do the answers to RQ1 & RQ2 depend on: **(a)** the frequency with which an entity-mention occurs in the data; and **(b)** the distribution of that entity across the news topics?"*) is that both the identification and use of entities, and the extent to which they are represented by their type each depend on frequency and topic distribution. Specifically, the frequency is depended on for the masked language modeling, and the topic distribution for the topic classification task.

6 Conclusion

We have presented RefNews-12, a novel news topic classification dataset. This dataset was collected by scraping Wikipedia articles for links. This collection method allows it to be expanded with additional topics and languages in the future. Because the Wikipedia pages also link to the pages of entities relevant to the incident, the dataset can be bridged easily to knowledge from Wikidata.

We have investigated entity representations in Transformer-based language models. We find that after having been fine-tuned for news topic classification these models do identify and use the entities to accomplish the task at hand. Although, whether they are used also depends on the task for which the model is trained. Our results also show that on average these language models do not represent entities only by their type. Entities are used by the model as distinctly different even within the same type. Finally, we have shown that the frequency with which an entity occurs in the data does not play a significant role in models performing topic classification. Nor does the topic distribution play a significant role in masked language modeling.

We obtained our results by altering the inputs of a model and measuring the change in performance. Crucially, to allow us to draw conclusions from these results we also control for model uncertainty. We believe this general methodology can be used to probe for many kinds of properties. As such it provides an additional probing technique which can be used to strengthen existing experimental evidence in the future.

7 Limitations & Future work

Our results are based on DistilBERT, which is a relatively small model. Because of this, the results are not necessarily representative of all Transformer-based language models. A further limitation is that our experiments are only performed with our RefNews-12 dataset, and the only downstream task we evaluate on is topic classification. Finally, the entity types we use are limited to the highest level of types (locations, organizations, persons and miscellaneous). It is possible that at more fine-grained levels entity representations do become less and less distinct.

In future work, we mean to address these limitations by including larger models and other datasets to show if the same patterns hold on all Transformer-based models including on other data and tasks. Also, by including entity linking in future experimentation, we will be able to extract entity-types from knowledge bases such as Wikidata, and perform substitutions exclusively within those much more fine-grained types.

Acknowledgments

This research was supported by Huawei Finland through the DreamsLab project. All content represented the opinions of the authors, which were not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks.

Russa Biswas, Radina Sofronova, Mehwish Alam, Nicolas Heist, Heiko Paulheim, and Harald Sack. 2021. Do Judge an Entity by Its Name! Entity Typing Using Language Models. In *The Semantic Web: ESWC 2021 Satellite Events*, Lecture Notes in Computer Science, pages 65–70, Cham. Springer International Publishing.

Samuel Broscheit. 2019. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685, Hong Kong, China. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. 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.

Fredrik K. Gustafsson, Martin Danelljan, and Thomas B. Schon. 2020. Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision. pages 318–319.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Analyzing Entities and Topics in News Articles Using Statistical Topic Models. In *Intelligence and Security Informatics*, Lecture Notes in Computer Science, pages 93–104, Berlin, Heidelberg. Springer.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Tara Safavi and Danai Koutra. 2021. Relational World Knowledge Representation in Contextual Language Models: A Review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1053–1067, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda. 2020. Probing for Referential Information in Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4177–4189, Online. Association for Computational Linguistics.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sara Veldhoen, Dieuwke Hupkes, and Willem H. Zuidema. 2016. Diagnostic Classifiers Revealing how Neural Networks Process Hierarchical Structure.

Piek Vossen, Filip Ilievski, Marten Postma, Antske Fokkens, Gosse Minnema, and Levi Remijnse. 2020. Large-scale Cross-lingual Language Resources for Referencing and Framing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3162–3171, Marseille, France. European Language Resources Association.

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. 2018. Don't Annotate, but Validate: a Data-to-Text Method for Capturing Event Data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wenzheng Zhang, Wenyue Hua, and Karl Stratos. 2021. EntQA: Entity Linking as Question Answering.