New or Old? Exploring How Pre-Trained Language Models Represent Discourse Entities

Anonymous ACL submission

Abstract

Recent research shows that pre-trained language models, built to generate text conditioned on some context, learn to encode syntactic knowledge to a certain degree. This has motivated researchers to move beyond the sentence-level and look into their ability to encode less studied discourse-level phenomena. In this paper, we add to the body of probing research by investigating discourse entity representations in large pre-trained language models in English. Motivated by early theories of discourse and key pieces of previous work, we focus on the information status of entities as discourse-new or discourse-old. We present two probing models, one based on binary classification and another one on sequence labeling. The results of our experiments show that pre-trained language models do encode information on whether an entity has been introduced before or not in the discourse. However, this information alone is not sufficient to find the entities in a discourse, opening up interesting questions about the definition of entities for future work.

1 Introduction

In a seminal paper from 1969, Karttunen imagines “a device designed to read a text in some natural language, interpret it, and store the content in some manner, say, for the purpose of being able to answer questions about it”. Such a device—considered by him “not a practical idea, for the time being at least”—he says would need to have a particular feature, namely that it “be able to recognize when a novel individual is mentioned in the input text and to store it along with its characterization for future reference.”

Now, more than 50 years later, neural models appear to have made such a device a practical idea after all. But do they recognize when a text introduces a new entity into the “universe of discourse”, or when, in contrast, the new information concerns a previously introduced one? This is the question that we are asking in this paper.

Following Karttunen’s idea, we focus here on discourse entities and in order to answer the question, we chose the task of distinguishing between the status of entity mentions (entity references) as discourse-new or discourse-old. Specifically, we build probing models that take as input the representations of pre-trained English language models and evaluate the models on this task.

We present two probing models tackling the task on two different levels of complexity: binary classification and sequence labeling. The first probe takes one entity mention and its preceding context up to that point, and produces a binary decision (discourse-new/-old). It tells us to what extent the context matters for this task. Inspired by the Named Entity Recognition (NER) task, the second probe labels each token in a sequence (new/old/outside). It tells us thus whether the entities can be localized in the sequence and labeled with the correct type. A few pieces of research have found first indications about the presence of entity knowledge in pre-trained language models. In particular, Sorodoc et al. (2020) focus on identifying pronoun-antecedent pairs, leaving the question open of whether models have a notion of entity spans. Li et al. (2021) work only with synthetic data, which is simple in nature with short sentences and few entities. Last, Gupta and Durrett (2019b) find that pre-trained language model’s representations are unable to trace explicit entity state changes in recipes and physical processes.

Considering that discourse entities are central to discourse theories and meaning, there is still a lot of ground to cover. We take a step back and look at entities being referred to over time. We believe that the new/old distinction lets us ask whether language models also keep track of the discourse entities.

Our findings suggest that contextualized pre-trained language model representations generated...
2 Related Work

The intuition that language models implicitly capture and in turn also benefit from entity knowledge has been explored for some time now (Ji et al., 2017; Yang et al., 2017), with recent papers focusing on how to inject some explicit entity representation into the system (Aina et al., 2019; Gupta and Durrett, 2019a, among others). The information status, which distinguishes between entities that are newly introduced in the discourse and those that are already known to the comprehender (Prince, 1992; Kamp and Reyle, 1993), has been studied in the context of bridging anaphora resolution, a non-identical relation between an anaphor and its antecedent (Markert et al., 2012; Hou, 2020, among others) based on semantic content.

Our work is most similar to that of Sorodoc et al. (2020) and Li et al. (2021). Both of these works are interested in probing entity knowledge in pre-trained language models at the discourse level, and they both take a semantic approach in that they are interested in the similarity between different mentions of the same entity at different points in the discourse.

Working with the OntoNotes (Pradhan et al., 2012) coreference corpus, Sorodoc et al. test whether pre-trained language model representations have the morpho-syntactic and semantic knowledge required to match a pronoun with its antecedent. They report results based on pre-trained representations generated with both a Transformer and an LSTM model. Using two baselines i) always referring to the nearest mention, ii) always referring to the most similar (cosine similarity) token, they found that a probe fed with the pre-trained embeddings succeeds at the task of predicting the correct antecedent (75.9% accuracy). An error analysis of the probe with the Transformer representations showed that noun phrases were harder to solve than pronouns (so they focus on the latter). The probe also succeeded in learning agreement, as tested by inserting distractors, but accuracy drops to 53% in hard cases (e.g., when the pronoun and antecedent disagree in gender/number). In our experiments, we go a step further and consider pronouns as well as noun phrase mentions.

Focusing on Transformers, Li et al., on their side, work with the Alchemy (derived from Long et al. (2016)) and Textworld (Côté et al., 2019) datasets. They use the data to construct logical propositions which are then classified into True/False with a binary classifier probe (e.g., You see an open chest. The only thing in the chest is an old key. The chest contains an apple. → True/False). This data transformation is possible because the original data are constructed short documents with simple sentences and few entities. It should be noted as well that although the probe itself is a low-capacity linear classifier, it needs a proposition embedder and a localizer of the entity in the sequence as additional pipeline components. Their results are measured through accuracy (the aggregation of all propositions with an entity) and they go as high as 94%. Our work is concerned with real world text data instead. As this increases the space of possible propositions, it also requires us to simplify the task, which we will explain in detail in the next sections.

Turning to the approach of probing or diagnostic classifiers (Hewitt and Liang, 2019), these are simple systems trained on the encoded representations of another system. If the probe succeeds in the task for which it is trained—discourse-new/discourse-old in our case, we conclude that the input had the necessary knowledge to solve the task.

Research based on probing models mostly relies on classification tasks. This paper is a part of that, but here we additionally use a sequence labeling task. This strategy has precedent in examples such as Ramponi et al. (2020) and Dai et al. (2019a) who use sequence labeling for event and entity extraction, respectively. In the context of probing discourse knowledge, Koto et al. (2021) has used this format for a sentence ranking experiment where the probe was asked to predict the most likely sentence ordering as a sequence.

3 Data and Pre-trained Model

We present two sets of probing experiments on the hidden representations of a pre-trained model using real-world text data.
(1) [Ripples from [the strike by [55,000 Machinists union members]_{new} against [Boeing Co.]_{new} reached air carriers Friday as [America West Airlines]_{new} announced [it]_{old} will postpone [its]_{new} new service out of Houston]_{new} because of [delays in receiving aircraft from [the Seattle jet maker]_{old}]. [Peter Otradovec, vice president for planning at the Phoenix, Ariz., carrier]_{new}, said in [an interview]_{new} that [the work stoppage at Boeing]_{old}, now entering its 13th day,_{old} “has caused [some turmoil in our scheduling]_{new}”.

(2) [Ripples from the strike by 55,000 Machinists union members against Boeing Co.] reached air carriers Friday as [America West Airlines] announced [it] will postpone [its new service out of Houston] because of [delays in receiving aircraft from the Seattle jet maker]. [Peter Otradovec, vice president for planning at the Phoenix, Ariz., carrier], said in [an interview] that [the work stoppage at Boeing, now entering its 13th day], “has caused [some turmoil in our scheduling]”.

Figure 1: Excerpt from one document extracted from the ARRAU corpus. (1) presents the original mentions with their new/old labels. (2) shows our formatting into IOB labels for the task of identifying and distinguishing between discourse-new and discourse-old entities (see Sec. 4.2).

### 3.1 Data

For our experiments, we use data from the ARRAU corpus (Uryupina et al., 2020) which is gold annotated with the information status of the mentions. We kept the official split used in the CRAC shared task (Poesio et al., 2018), but discarded the dialog and narrative genres (trains and pears, respectively) because they are presumably different from the domain of the pre-trained model and are not publicly available. We construct two different variants for the probing tasks, spans and heads. The examples per partition can be found in Table 1.

#### Spans

As Fig. 1 shows, the original ARRAU data contains several embedded entities (1). The sequence labeling task (cf. Sec. 4.2) requires some adaption of the data. In order to use the same data for both tasks, we perform the same pre-processing: We remove all embedded entities from the maximal entity span containing them, but keep their re-mentions. For example, the mention the strike by 55,000 Machinists union members against Boeing Co. has been removed from the outer entity span Ripples from the strike ..., but the re-mention the work stoppage at Boeing, now entering its 13th day is kept (2). We expect this version of the data to be easier for the classification probe as it has access to all the information in the span, but harder for the sequence labeling probe, since there is much variability in the length of the mentions. Identifying the span also implies higher order knowledge of the notion of entity. Neural coreference resolution systems, for instance, explicitly learn the representation of spans as the combination of the vectors for the first and last tokens and the head of a mention (Lee et al., 2017).

#### Heads

In this version, we take only the head of each mention. This means that the probing model only needs to label the lexical head or pronoun, as opposed to the complete span of a mention. We obtained the heads by extracting the min_words attribute from the manual ARRAU annotations.² Here, embedded entities are retained.

### 3.2 Pre-trained Representations

All of our probes are based on the representations learned by a pre-trained Transformer-XL model (Dai et al., 2019b) (available through the Hugging Face library as TRANSFO-XL-WT103). We focus on this English model exclusively as it has been

²Annotators were not limited to one word, so a head can still contain more than one token, as in Resolution Trust Corp.
used in the closely related work by Sorodoc et al. (2020), but our code is easily extensible to work with representations from other language models as well. The scope of our claims is thus limited to this pre-trained model. The Transformer-XL model is explicitly able to capture long contexts and generate “relatively” coherent long texts (Dai et al., 2019b) by using a recurrence mechanism over cached previous segment states. This counters the “context fragmentation” introduced by chopping off contexts at a given length to cope with limited computational capacities.

We extract the last 1024-dimensional hidden state representations for each token by feeding the pre-trained model a whole document at a time, so they are contextualized with the discourse knowledge encoded by the model.

3.2.1 Baselines
To interpret the results of our probing models, we compare them to models initialized with static 300-dimensional fastText embeddings (Grave et al., 2018), which we extract word by word using the Python fastText module.\footnote{https://fasttext.cc/docs/en/python-module.html} We further match the results against two simple baselines, a majority baseline that labels every entity as new, and one based on POS tags where only entity mentions that start with a definite article or pronouns are considered old. The intuition behind the second one is that these are easy and frequent cases, which may reveal whether our models simply rely on these linguistic cues.

4 Probing Experiments
We first perform a classification task to gather information about the entity representations themselves. The second probing task looks at sequence labeling in order to evaluate whether this information can be also used to detect entity boundaries.

4.1 Classification Task
Previous work on entity status tracking (Gupta and Durrett, 2019b) framed the task as entity classification by pre-extracting the entity in question, thereby not requiring the model to identify what an entity is in the first place. We adapt this task in order to probe whether pre-trained hidden representations contain information about whether an entity is newly introduced into the discourse or if it is a re-mention of an already introduced entity.

4.1.1 Pre-processing
First, we split each text incrementally at each entity mention such that the context contains the n words up to the first entity (and in the next sample the second and so on), and the target contains the respective following entity tokens. As described in Sec. 3.1 we extract either only the heads or the maximal spans of entities as the target. We prepend the {eos} token to the context to avoid empty contexts when the first token is (part of) an entity (e.g., in the very first sentence of a document).

Next, we use the pre-trained Transformer-XL representations for every context token, and sum over the extracted target entity vectors to get an entity representation (see lower part of Fig. 2).

4.1.2 Models
Following the baseline models of Gupta and Durrett (2019b), we train an attention-based classifier that computes bilinear attention (Luong et al., 2015) between the entity representation and the context tokens and predicts the entity category (new/old) from the combined result. We use the Attention implementation from pytorch-nlp (Petrochuk, 2018) (Eq. 1-4) and a linear layer with a sigmoid function on top (Eq. 5).

\[
\begin{align*}
 a_i &= h_{c_i}^T \ast W_{att} \ast e & (1) \\
 \alpha &= \text{softmax}(a) & (2) \\
 \text{context} &= \sum \alpha_i \ast h_{c_i} & (3) \\
 h_{c,e} &= \tanh(W_{comb} \ast [\text{context}, e]) & (4) \\
 P(y|h_{c1}, \ldots, h_{cn}, e) &= \text{sigmoid}(W_a \ast h_{c,e} + b) & (5)
\end{align*}
\]

This model has access to the whole context up to the entity. To gather further insights on the role of the context tokens in terms of what kind of information is already encoded in the entity representation itself, we additionally train a model without context, using only the entity representation to predict its status (Eq. 6).

\[
P(y|e) = \text{sigmoid}(W_b \ast e + b) & (6)
\]

The model architectures are displayed in Fig. 2. All models are trained with the BCEWithLogitsLoss and use the Adam optimizer with a learning rate of 0.001 and a batch size of 64. Early stopping is applied based on the loss on the development set.
The results of the classification experiments are displayed in Table 2. While the fastText embeddings already yield an improvement over the majority and the POS baseline, the Transformer-XL embeddings yield the best overall results, suggesting that the contextualization adds some useful information on the entity state. Surprisingly, however, there is not much difference between the attention-based and the entity-based models, suggesting that the information required for this task is contained in the pre-trained representations themselves. It is also interesting that taking the whole span into account improves the results for the fasttext embeddings, but yields no gain for the Transformer-XL representations. This suggests that due to the contextualization, the necessary information is already encoded in the head representation itself. A detailed discussion follows in Sec. 5.

4.2 Sequence Labeling Task

Inspired by the NER scenario, our second probing model takes the form of a sequence labeling task, whereby each token in the sequence is assigned a categorical label. Discourse entities are a broader category of named entities, so instead of assigning entities a type (e.g., ORGANIZATION, PERSON, TIME, etc.), the probe assigns new or old labels. Concretely, following the IOB scheme, an entity’s first token is either B-NEW or B-OLD and subsequent tokens have an I-NEW or I-OLD label. Non-entity tokens are labeled as outside (O). An example is shown in Fig. 1.

This framework offers us two advantages in a single task: i) the probe has to localize the entity in the sequence (an additional pre-processing step in previous work), and ii) the probe has to assign a classification label to the entity.

4.2.1 Pre-processing

We slice the extracted hidden vector sequence according to the tokens in the original sentences, in order to feed our probes with examples at the sentence-level. In other words—and taking the example in Fig. 1—the hidden representations are extracted based on the whole document, but the probing model labels the document sentence by sentence. Because this precludes the probe from accessing any embeddings beyond the sentence, any success at predicting discourse-new or discourse-old should come from the entity embeddings themselves. This has the added advantage of easing the computation cost.

4.2.2 Model

The underlying method for our probe is a linear chain conditional random field (CRF) model. The input to the model are sequences of n vectors, where each input \( h_i \) is a contextualized vector with size 1024 yielded by the pre-trained language model. This sequence of pre-trained vectors is fed into an LSTM layer (Eq. 7) and a rectified linear unit activation function (Eq. 8), before being resized (Eq. 9) in order to fit the CRF (Eq. 10). The CRF layer finds the best possible sequence of labels \( (y_1, ..., y_n) \) for the entire input sequence.

\[
\begin{align*}
    r_i &= LSTM(h_i, r_{i-1}) \\
    o_i &= RELU(r_i) \\
    l_i &= W \ast o_i + b \\
    p(y_1, ..., y_n| h_1, ..., h_n) &= CRF(l_1, ..., l_n)
\end{align*}
\]

We present experiments with and without the LSTM layer, the Transformer-XL architecture, and the pre-trained representations themselves. This has the added advantage of easing the computation cost.

5We took advantage of the freely available implementation at https://github.com/kmkurn/pytorch-crf
We build several model versions to estimate the Transformer-XL model. After the LSTM layer (Eq. 7-8), a linear layer (Eq. 9) is needed to reduce the dimensions of the LSTM output from the hidden size to label size required by the CRF (Eq. 10). The CRF has a choice among 5 labels at each time step. The CRF has a high deviation is between 0.00 and 0.04 for all versions), and a POS-based (pronouns and defNP = discourse-old) and majority class (discourse-new) baseline comes after a B), and the LSTM learns the semantic content (i.e., new vs old).

The training takes between 1 and 2 epochs using early stopping based on the loss computed on the development set. We use mini-batching with size 64, a hidden size of 256 for the LSTM output, and a learning rate of 0.01 with an Adam optimizer. Dropout of 0.2 is also applied.

### 4.2.3 Baselines

We build several model versions to estimate the success of our probes: In addition to the majority and POS baselines (cf. Sec. 3.2.1), we build a simple CRF using the Scikit-learn (Pedregosa et al., 2011) compatible CRFsuite (Okazaki, 2007) wrapper (Korobov, 2015) based on simple surface form features. These include whether a token is at the beginning or end of a sequence, whether the token starts with a capital letter, the last three characters of the token, and the last two characters of a token. We also add two versions of our probing models that do not rely on pre-trained representations but train the embeddings from scratch with dimensions 1024 for the comparison to Transformer-XL and 300 for fastText.

### 4.2.4 Results

The results for all configurations are reported in Table 3. For computing the scores, we used the SeqEval package (Ramshaw and Marcus, 1995; Nakayama, 2018).

A very clear pattern emerges: Heads are easier to identify than Spans, and discourse-old is easier to predict than discourse-new. It follows that the combination of span + discourse-new is the most difficult category, and one in which the probes are surpassed by the baselines. Besides, the models with the LSTM yield consistently higher results than the models relying on the CRF only.

When considering the type of input, there is a similar pattern to the classifier. Although the results using static embeddings are better than the baselines, the contextualized Transformer-XL representations present a systematic improvement overall. Interestingly, this improvement is more marked for the Spans and negligible for Heads in additional experiments with a CRF for entity identification only, i.e., without labeling the entities as new or old (cf. Appendix B, Table 5). This keeps with the intuition that identifying the heads is akin to finding nouns, but identifying the relevant spans which are also entities is more complex, involving discourse-level knowledge.
This model is superior in the Heads version of the discourse-new class. We believe this is because the binary classifier does slightly better with the Spans setting (comparing F1 scores), but in the sequence labeling experiments, in particular of the entities, discourse-old seems easier to predict overall results than any of our probing models. The high success of the binary classifier probe confirms that classifying an entity as new or old is not challenging, provided that the model has access to the entity representations. Comparable to a coreference resolution system, this model is fed with an aggregated representation comprising all tokens in the mention, so it does not need to locate the entity in the sequence. Indeed, finding the entities given a sequence is the hard part in our task, as confirmed by the sequence labeling probe. This model is superior in the Heads version of the data, but less successful with Spans, where the boundaries of each entity must be identified. In this task, the simple POS-based baseline yields better overall results than any of our probing models.

**Error Analysis:** Looking at the length of mentions (Fig. 4), interestingly, the systems seem more error prone only when the number of tokens in a mention exceeds a high threshold (8 for Heads and 15 for Spans). Despite being the minority class with about 40% of the entities, discourse-old seems easier to predict in the sequence labeling experiments, in particular for the Spans setting (comparing F1 scores), but the binary classifier does slightly better with the discourse-new class. We believe this is because the classifier is given the whole entity representations (the sum of tokens comprising each entity), and they contain the new/old information. The CRF, however, does entity identification and labeling jointly, intrinsically a harder task, in particular for the discourse-new class. It is worth stressing that, although the pre-trained representations were extracted at the document-level, only the sentence-level examples are fed to the CRF. Thus, whatever discourse knowledge the probes pick up must come from the pre-trained embeddings conforming the entity itself. This conclusion is backed by the classifier results, which are extremely close between experiments, regardless of whether context is used. Pronouns and definite noun phrases are two types of mentions with enough regularity (i.e.,

<table>
<thead>
<tr>
<th></th>
<th>Discourse New</th>
<th>Discourse Old</th>
<th>Avg.F1</th>
<th>Discourse New</th>
<th>Discourse Old</th>
<th>Avg.F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
<td>F1</td>
<td>Prec</td>
<td>Rec</td>
<td>F1</td>
</tr>
<tr>
<td><strong>Probing Transformer-XL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM + Linear + CRF</td>
<td>0.75</td>
<td>0.79</td>
<td>0.77</td>
<td>0.80</td>
<td>0.78</td>
<td>0.79</td>
</tr>
<tr>
<td>Linear + CRF</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.75</td>
<td>0.69</td>
<td>0.72</td>
</tr>
<tr>
<td>LSTM + Linear + CRF_\text{scratch}</td>
<td>0.59</td>
<td>0.71</td>
<td>0.64</td>
<td>0.74</td>
<td>0.54</td>
<td>0.62</td>
</tr>
<tr>
<td>Linear + CRF_\text{scratch}</td>
<td>0.51</td>
<td>0.59</td>
<td>0.55</td>
<td>0.63</td>
<td>0.47</td>
<td>0.53</td>
</tr>
<tr>
<td><strong>Baselines fastText 300</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM + Linear + CRF</td>
<td>0.67</td>
<td>0.76</td>
<td>0.71</td>
<td>0.75</td>
<td>0.63</td>
<td>0.68</td>
</tr>
<tr>
<td>Linear + CRF</td>
<td>0.55</td>
<td>0.63</td>
<td>0.59</td>
<td>0.69</td>
<td>0.45</td>
<td>0.55</td>
</tr>
<tr>
<td>LSTM + Linear + CRF_\text{scratch}</td>
<td>0.59</td>
<td>0.70</td>
<td>0.64</td>
<td>0.72</td>
<td>0.54</td>
<td>0.62</td>
</tr>
<tr>
<td>Linear + CRF_\text{scratch}</td>
<td>0.53</td>
<td>0.62</td>
<td>0.57</td>
<td>0.65</td>
<td>0.46</td>
<td>0.53</td>
</tr>
<tr>
<td><strong>Baselines w/o embeddings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple CRF</td>
<td>0.57</td>
<td>0.70</td>
<td>0.62</td>
<td>0.71</td>
<td>0.45</td>
<td>0.55</td>
</tr>
<tr>
<td>POS baseline</td>
<td>0.65</td>
<td>0.51</td>
<td>0.57</td>
<td>0.51</td>
<td>0.58</td>
<td>0.55</td>
</tr>
<tr>
<td>Majority class</td>
<td>0.50</td>
<td>1.00</td>
<td>0.74</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 3: Average results from five different random seeds for all discourse-new vs. discourse-old sequence labeling models, probing pre-trained Transformer-XL representations vs. static fasttext embeddings and embeddings trained from scratch (standard deviation is between 0.01 and 0.06 for all versions). Baselines also include a simple CRF with surface features, a POS-based (pronouns and defNP = discourse-old) and the majority class (discourse-new) baseline.

## 5 Discussion and Analysis

Our results confirm previous findings about the presence of entity knowledge in pre-trained language models. We find that the entity knowledge extends to different types of entity mentions extracted from natural data.

The high success of the binary classifier probe confirms that classifying an entity as new or old is not challenging, provided that the model has access to the entity representations. Comparable to a coreference resolution system, this model is fed with an aggregated representation comprising all tokens in the mention, so it does not need to locate the entity in the sequence. Indeed, finding the entities given a sequence is the hard part in our task, as confirmed by the sequence labeling probe. This model is superior in the Heads version of the data, but less successful with Spans, where the boundaries of each entity must be identified. In this task, the simple POS-based baseline yields better overall results than any of our probing models.

**Error Analysis:** Looking at the length of mentions (Fig. 4), interestingly, the systems seem more error prone only when the number of tokens in a mention exceeds a high threshold (8 for Heads and 15 for Spans). Despite being the minority class with about 40% of the entities, discourse-old seems easier to predict in the sequence labeling experiments, in particular for the Spans setting (comparing F1 scores), but the binary classifier does slightly better with the discourse-new class. We believe this is because the classifier is given the whole entity representations (the sum of tokens comprising each entity), and they contain the new/old information. The CRF, however, does entity identification and labeling jointly, intrinsically a harder task, in particular for the discourse-new class. It is worth stressing that, although the pre-trained representations were extracted at the document-level, only the sentence-level examples are fed to the CRF. Thus, whatever discourse knowledge the probes pick up must come from the pre-trained embeddings conforming the entity itself. This conclusion is backed by the classifier results, which are extremely close between experiments, regardless of whether context is used. Pronouns and definite noun phrases are two types of mentions with enough regularity (i.e.,

Figure 4: % of errors per mention length in test data.
When we look at the less powerful probing models (Heads and Spans, respectively). This shows that, Transformer-XL sequence labeling probes compared to the POS baseline. The parentheses enclose the number of pronouns.

<table>
<thead>
<tr>
<th>Label</th>
<th>Boundaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF\text{heads}</td>
<td>2 (3) 137 (29)</td>
</tr>
<tr>
<td>CRF + LSTM\text{heads}</td>
<td>10 (4) 35 (13)</td>
</tr>
<tr>
<td>CRF\text{spans}</td>
<td>82 (4) 106 (23)</td>
</tr>
<tr>
<td>CRF + LSTM\text{spans}</td>
<td>83 (3) 65 (14)</td>
</tr>
</tbody>
</table>

Table 4: Number of error types produced by the Transformer-XL sequence labeling probes compared to the POS baseline. The parentheses enclose the number of pronouns.

The set of forms and determiner the) for the models to exploit frequency heuristics. In this sense, comparing with the POS baseline constitutes an interesting case study (Table 4). The fact that these two make different predictions is a sign that our probes are not deterministically exploiting this heuristic. Table 4 also confirms that finding an entity’s boundaries is harder than determining its label. Inspecting the differences closely shows that the errors correspond to the pronouns it, this, that and which, known to be problematic for coreference resolution. For the classifier, we also found that it and that are amongst the most common errors (18/424 in spans, 23/753 in heads). We also inspected the definite noun phrases, but could not identify any specific pattern in the errors.

**Model Analysis:** There is no benefit in using static embeddings (Linear + CRF fastText) vs simple classic features (simple CRF) for this task. Comparing the pre-trained representations with the models that were trained from scratch, we observe the following: i) For the most powerful model (the ones with the LSTM layer) the gain of using pre-trained embeddings is 0.15 and 0.19 in average F1 for Transformer-XL and 0.7 and 0.8 for fastText (Heads and Spans, respectively). This shows that, while the embedding size does not have an impact if embeddings are trained from scratch (similar results for 1024 and 300 dimensions), the contextualized Transformer-XL representations contain more useful information for the probing task. ii) When we look at the less powerful probing models (without the LSTM), the differences are 0.17 and 0.16 in average F1 for Transformer-XL, and 0.2 and -0.3 for fastText, showing that the LSTM is necessary to extract any useful information from the static embeddings. Transformer-XL embeddings, on the other hand, already benefit from the contextualization during pre-training, as we see similar improvements as those obtained with the complex model.

The LSTM models do yield better scores overall, suggesting that additional contextualization on the sentence level helps for this task. Collectively, these results signal that document-level contextualization does help to encode the new/old distinction, but not as much as one might have expected. If a model is presented with an entity, determining its status is not hard, even without contextualization. However, finding an entity in a sentence or discourse is challenging, even for powerful Transformer-XL representations. This raises the question of whether pre-trained language models are able to identify entities in the wild.

### 6 Conclusions and Future Work

We have built two probing models for the task of identifying the discourse status of entities as new or old. Our models use binary classification and sequence labeling with input representations from a Transformer-XL language model. Not only confirming, but also advancing the findings from previous work, we have found that the pre-trained representations do encode the old/new information in the tokens comprising the entity. However, localizing the entity boundaries within the sentence is difficult for the sequence labeling probe. This leaves interesting questions for future work, in particular, defining what an entity is and what it looks like, as understanding entities is a step towards understanding discourse and meaning. In this work, we have limited ourselves to entities appearing in co-referential chains, meaning that we have excluded all mentions annotated as non-referring. In light of our findings, we believe that distinguishing between entities and non-entities is the next challenge. In this sense, one could imagine a task where a probe is asked to differentiate between referring and non-referring mentions, a hard problem in the context of coreference resolution.

Pre-trained language models are common resources in our community. They are widely used in pipelines for many applications and also as objects of linguistic study. It is therefore crucial that we continue our efforts towards understanding their abilities and limitations. We now know that they are capable of encoding syntactic knowledge (Schrimpf et al., 2020), discourse seems then, the next stage. In this light, it seems that the time for the device envisioned by Karttunen (1969) has (sort of!) come.
7 Ethical Considerations and Limitations

The models trained in this study are not optimized to solve specific tasks in the best possible way, but to gain insights about the underlying representations and thus the abilities of pre-trained language models, which are sometimes attributed human like language-generation abilities. However, all findings are only applicable to the model under investigation (Transformer-XL) and any claims are specific to English. Reproducing our work requires access to the ARRAU corpus, on which we base all of our experiments.

References


**A Probing experiment details**

We trained all models on an NVIDIA GeForce GTX 1080 Ti. Each classification experiment took 10 to 15 minutes and each sequence labeling experiment took between 7 and 20 minutes.

**B Sequence labeling results without new/old labels**
<table>
<thead>
<tr>
<th></th>
<th>Heads</th>
<th>Spans</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec.</td>
<td>Rec.</td>
</tr>
<tr>
<td><strong>Probing Transformer-XL</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM + Linear + CRF</td>
<td>0.86</td>
<td>0.91</td>
</tr>
<tr>
<td>Linear + CRF</td>
<td>0.84</td>
<td>0.79</td>
</tr>
<tr>
<td><strong>Baselines fastText 300</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM + Linear + CRF</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td>Linear + CRF</td>
<td>0.75</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 5: Single-run results from sequence labeling experiments for entity identification without predicting their status as new or old.