Exploring Span Representations in Neural Coreference Resolution

Patrick Kahardipraja[†], **Olena Vyshnevska**[†], **Sharid Loáiciga**
Computational Linguistics, Department of Linguistics
University of Potsdam, Germany
{kahardipraja,olena.vyshnevska,loaicigasanchez}@uni-potsdam.de

Abstract

In coreference resolution, span representations play a key role to predict coreference links accurately. We present a thorough examination of the span representation derived by applying BERT on coreference resolution (Joshi et al., 2019) using a probing model. Our results show that the span representation is able to encode a significant amount of coreference information. In addition, we find that the head-finding attention mechanism involved in creating the spans is crucial in encoding coreference knowledge. Last, our analysis shows that the span representation cannot capture non-local coreference as efficiently as local coreference.

1 Introduction

Coreference resolution, the task of grouping all referring expressions that point to the same entity into a cluster, plays a key role for various higher level NLP tasks that involve natural language understanding such as information extraction, question answering, machine translation, text summarisation, and textual entailment. Referring expressions or mentions can be common nouns, proper nouns, or pronouns, which refer to a real-world entity known as the referent.

With the breakthrough of end-to-end neural systems (Lee et al., 2017), current coreference resolution systems are for the most part neural based. Contrary to previous architectures which identified mentions and then took coreferential decisions in two separate steps, these systems jointly learn the two. A typical system requires different levels of semantic representation of the input sentences, usually done by computing representations at the span level given the word embeddings.

In another area, a wave of recent work has tried to inspect neural NLP models by associating neural network components with distinct linguistic phenomena by means of probing tasks (Shi et al., 2016; Liu et al., 2019a; Tenney et al., 2019).

Targeting the coreference task, in this paper, we build a probing model (Tenney et al., 2019; Liu et al., 2019a) to find out what degree of coreference information is encoded in the span representations as first proposed by Lee et al. (2017). Specifically, we generate mention-span representations with BERT embeddings fine-tuned on the OntoNotes dataset (Pradhan et al., 2012) and train a probing model to predict coreference arcs between two mentions from the mention-span representations alone. Moreover, we explore how fine-tuning BERT (Devlin et al., 2019) on coreference resolution affects the linguistic knowledge learned by the span representations. Given the well-documented difficulty in modelling long-distance coreference relations, we also measure the robustness of the span representations at different distance ranges between mentions.

Our probing models consistently achieve $> 90\%$ accuracy and F1, suggesting that span representations encode a significant amount of coreference information. Besides, they show that fine-tuning a BERT model greatly helps with encoding coreference relations. By ablating components of the span representation, we also find that the head-finding attention mechanism plays a crucial part in encoding important coreference information. Finally, we show that despite using a fine-tuned BERT, the span representations cannot capture non-local coreference relation efficiently. Our implementation is publicly available[1].

[†] Shared first authorship

[1] https://github.com/pkhdipraja/exploring-span-representations

Proceedings of the First Workshop on Computational Approaches to Discourse, pages 32–41
Online, November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

2 Related Work

2.1 Span-Ranking Architecture

In this paper we focus on the span representation used in span-ranking models (Lee et al., 2017, 2018; Joshi et al., 2019) and examine their capability to encode the necessary information to make coreference decisions.

Lee et al. (2017) proposed an end-to-end coreference resolution model that learns to jointly model mention detection and coreference prediction using span-ranking. However, the model only computes scores between pairs of entity mentions. In an attempt to improve the weakness of this approach, Lee et al. (2018) proposed a model that captures higher-order interactions between mention spans in predicted coreference clusters. The model refines existing span representations iteratively with the antecedent distribution as an attention mechanism. We further refer to this model as *c2f-coref*.

Joshi et al. (2019) proposed to replace the bidirectional LSTM encoder in *c2f-coref* with BERT transformers and fine-tune it for coreference resolution. Although BERT improves the state-of-the-art results in other NLP tasks significantly (Devlin et al., 2019), coreference resolution still proves to be a challenging task, as the BERT encoder offers a marginal performance increase only. Furthermore, the model still struggles in modelling pronouns and resolving cases where mention paraphrasing is required. We further refer to this model as *BERT-coref*.

2.2 Probing Tasks

The most common method to explore linguistic properties in neural network components is by using the hidden state activations to predict the property of interest, also known as “probing tasks” (Conneau et al., 2018) or “auxiliary prediction tasks” (Adi et al., 2016). Shi et al. (2016) use the internal representations of an LSTM encoder as input to train a logistic regression classifier that predicts various syntactic properties. Conneau et al. (2018) study the linguistic properties of fixed-length sentence encoders with a bidirectional LSTM and gated convolutional networks.

Liu et al. (2019a) explore representations produced by pre-trained contextualisers and demonstrate that frozen contextual representations fed into linear models can show similar levels of performance as state-of-the-art task-specific models on many NLP tasks. They also used the coreference arc prediction task, whereby linear models are used to predict whether two mentions corefer. The coreference arc prediction was already used by Soon et al. (2001) as a part of the mention-pair model, where it is used with heuristic procedures to merge coreference chains.

Tenney et al. (2019), on their part, introduced the edge probing framework, which focuses on linguistic analysis on sub-sentence level. Their approach relies on a FFNN model with a projection layer and an attention mechanism on top of frozen contextual vectors to predict linguistic properties. Clark et al. (2019) further extended the probing-based approach by proposing attention-based probing classifiers and show that the attention heads in BERT correspond to linguistic notions of syntax and coreference.

Our approach is most similar to Liu et al. (2019a) and Tenney et al. (2019), but we use the span representation learned from Lee et al.’s 2017 coreference resolution model and focus on examining coreference phenomena. Note that we use the coreference arc prediction task as a tool to understand the span representation better, we do not do coreference resolution. Compared to Liu et al. (2019a) who consider single-token mentions only, we use mention-spans to predict coreference arcs. We also compare the span representation against a baseline span representation obtained from pre-trained contextual word embeddings (Tenney et al., 2019).

3 Probing Mention-Span Representations

3.1 Span Representations

Span representations are key in span-ranking models since they are used to compute a distribution over candidate antecedent spans. In order to predict coreference relations accurately, a span representation should also capture information about the span’s internal structure and its surrounding context. For our experiments, we construct span representations as proposed by Lee et al. (2017), but with BERT embeddings (Devlin et al., 2019) instead of an LSTM-based encoder to encode the lexical information of a span and its context, following Joshi et al. (2019). A span representation is a vector embedding which consists of context-dependent boundary representations with an attentional representation of the head words over the span. The boundary representations are composed of the first and last wordpieces of the span itself.

The head words are automatically learned using additive attention (Bahdanau et al., 2015) over each wordpiece in a span:

$$\alpha_t = \boldsymbol{w}_\alpha \cdot \mathrm{FFNN}_\alpha(\boldsymbol{x}_t^*)$$

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum\limits_{k=start(i)}^{end(i)} \exp(\alpha_k)}$$

$$\hat{\boldsymbol{x}}_i = \sum_{t=start(i)}^{end(i)} a_{i,t} \cdot \boldsymbol{x}_t$$

where $\hat{\boldsymbol{x}}_i$ is a weighted vector representation of wordpieces for span i. This representation is augmented by a $\mathbb{R}^d$ feature vector which encodes the size of span i with $d = 20$. The final representation $\boldsymbol{g}_i$ for span i is formulated as follows:

$$\boldsymbol{g}_i = [\boldsymbol{x}^*_{start(i)}, \boldsymbol{x}^*_{end(i)}, \hat{\boldsymbol{x}}_i, \phi_i]$$

where $\boldsymbol{x}^*_{start(i)}$ and $\boldsymbol{x}^*_{end(i)}$ are first and last wordpieces of a span, and ϕ_i is the span width embedding.

3.2 Coreference Arc Prediction

We focus on the coreference arc prediction task, which is a part of the probing tasks suite for contextual word embeddings. In this task, a probing model is trained to determine whether two mentions refer to the same entity. We produce negative samples following the approach by Liu et al. (2019a). For every pair of gold mentions (w_i, w_j), where they belong to the same gold coreference cluster and w_i is an antecedent of w_j, we generate a negative example (w_{random}, w_j) where w_{random} is randomly sampled from a different coreference cluster.

This method ensures a balanced ratio between positive and negative examples. The negative examples do not contain any singleton mentions, as in OntoNotes only coreferential mentions are annotated. We also follow the approach of Tenney et al. (2019) by using spans of wordpieces for mentions, as Liu et al.'s approach is limited to single-token mentions and therefore unable to fully exploit available information in a mention-span.

3.3 The Probing Model

Our probing model is a simple feed-forward neural network (FFNN), which is designed with a limited capacity to focus on the information that can be extracted from the span representations. As input to the model, we take a span representation for a pair of mention-spans $\boldsymbol{g}_1 = [\boldsymbol{x}^*_{start(1)}, \boldsymbol{x}^*_{end(1)}, \hat{\boldsymbol{x}}_1, \phi_1]$ and $\boldsymbol{g}_2 = [\boldsymbol{x}^*_{start(2)}, \boldsymbol{x}^*_{end(2)}, \hat{\boldsymbol{x}}_2, \phi_2]$, where both $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ are concatenated and passed to the FFNN. The FFNN consists of a single hidden layer followed by a sigmoid output layer. The model is trained to minimise binary cross-entropy with respect to the gold label $Y \in \{0, 1\}$. The probing architecture is depicted in Figure 1.

We obtained the mention-span representations from BERT, a language representation model based on the Transformer architecture (Vaswani et al., 2017), trained jointly with a masked language model and next sentence prediction objective. It enables significant improvement in many downstream tasks with relatively minimal task-specific fine-tuning. To study the quality of mention-span representations, we extract mention-span embeddings from BERT-base (12-layer Transformers, 768-hidden) and BERT-large (24-layer Transformers, 1024-hidden) pre-trained models. Furthermore, we compare these *original* BERT models with *fine-tuned* variants, with the purpose to assess any fine-tuning effect on the quality of the span representations.

4 Experiments

4.1 Dataset

We use the coreference resolution annotation from the CoNLL-2012 shared task based on the OntoNotes dataset (Pradhan et al., 2012). The dataset is split into 2,802 training documents, 343 validation documents, and 348 test documents. On average, the training documents contain 454 words. The largest document contains a maximum of 4,009 words. Since OntoNotes only provides annotations for positive examples, we generate our own negative examples (§3.2).

4.2 Implementation Details and Hyperparameters

We extend the original Tensorflow implementation of *BERT-coref*[2] in order to build our probing model with Keras frontend (Chollet et al., 2015). Our probing model is trained for 50 epochs, using early stopping with patience of 3 and batch size of 512. For optimisation, we use Adam (Kingma and Ba, 2015) with a learning rate of 0.001. The weights

[2] https://github.com/mandarjoshi90/coref

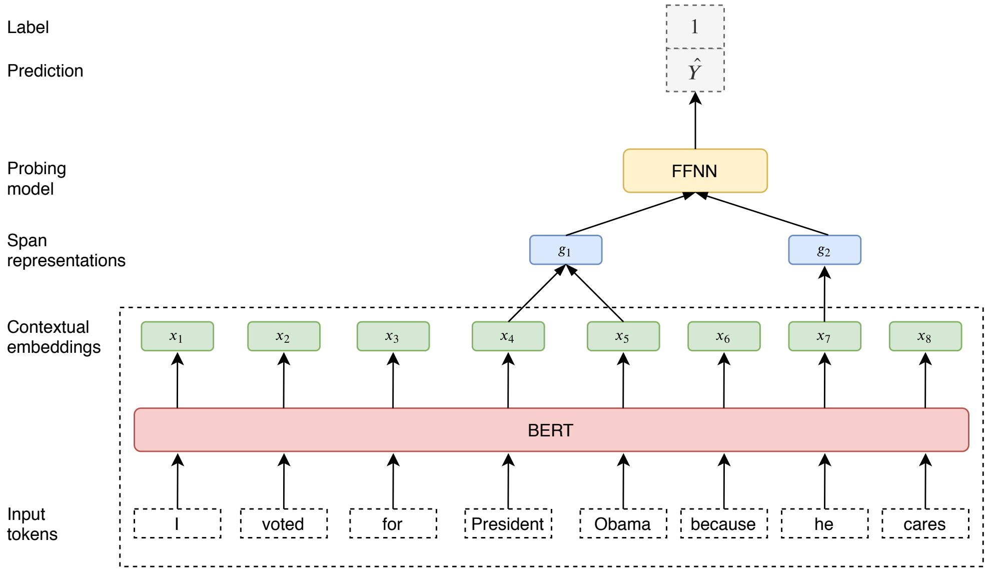


Figure 1: The probing architecture for span representations. The feed-forward neural network is trained to extract information from span representations $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$, while all the parameters inside the dashed line are frozen. The example depicts a mention-pair, where $\boldsymbol{g}_1$ corresponds to span representation of ”President Obama”, while $\boldsymbol{g}_2$ corresponds to ”he”. We predict $\hat{Y}$ as positive for this example.

of the probing model are initialised with Kaiming initialisation (He et al., 2015) and the size of the hidden layer is $d = 1024$ with rectified linear units (Nair and Hinton, 2010). As mentioned previously, we use both a pre-trained BERT (original) model without fine-tuning the encoder weights and a BERT model that has been fine-tuned on the coreference resolution task (i.e., on OntoNotes annotations). For the fine-tuned BERT model, we take the models that yield the best performance for Joshi et al. (2019), which were trained using 128 wordpieces for BERT-base and 384 wordpieces for BERT-large. The fine-tuned model is trained using split OntoNotes documents where each segment non-overlaps and is fed as a separate instance. This is done as BERT can only accept sequences of at most 512 wordpieces and typically OntoNotes documents require multiple segments to be read entirely. In all of our experiments, we use the cased English BERT models. We will further refer to the base and large variants as *BERT-base c2f* and *BERT-large c2f* respectively.

4.3 Baseline

As our baseline, we use the span representation introduced in the edge probing framework (Tenney et al., 2019). First of all, we take concatenated contextual embeddings for a pair of mention-spans $e^{(1)} = [x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, ..., x_n^{(1)}]$ and $e^{(2)} = [x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, ..., x_n^{(2)}]$ as inputs. We then project the concatenated contextual embeddings $e^{(1)}$ and $e^{(2)}$ to improve performance following Tenney et al. (2019):

$$e^{(i)} = Ae^{(i)} + b$$

where $i = (1, 2)$, A and b are weights of the projection layer. Afterwards, we apply the self-attentional pooling operator in §3.1 over the projected representations to yield fixed-length span representations.

This helps to model head words for each mention-span. These mention-span representations are then concatenated and passed to the probing model to predict whether they corefer or not. We use shared weights for both projection and self-attentional layer so that the model can learn the similarity between representations of mention-spans. It is important to note that the self-attention pooling is computed only using tokens within the boundary of the span. As a result, the model can only access information about the context surrounding the mention-span through the contextual embeddings. We take the contextual embeddings from activations of the original pre-trained BERT final layer, while freezing the encoder.

We compare the span representation used in the span-ranking model against the baseline, as it measures the performance that the probing model can achieve with representations that are constructed from lexical priors alone, without any access to the local context within the mention-spans. The resulting baseline span representation have a dimension of $d = 768$ for BERT-base and $d = 1024$ for BERT-large.

4.4 Long-range Coreference

In order to investigate whether the span representation is able to capture long-range coreference relations, we extend our baseline by introducing a convolutional layer to incorporate surrounding context and improve the baseline span representation, following Tenney et al. (2019).

We replace the projection layer in our probing architecture with a fully-connected 1D CNN layer with a kernel width of 3 and 5, stride of 1 and same padding to properly include contextual embeddings at the beginning and at the end of each mention-span. This is equivalent to seeing ± 1 and ± 3 tokens around the centre word respectively. We also initialise the weights of the CNN layer with Kaiming initialisation (He et al., 2015). Using this extended probing architecture with a CNN layer as another baseline, which we will refer to as *CNN-baseline*, enables us to examine the contribution of local and non-local context to the performance of the probing model.

We then test our probing model with various distances between mention-spans. We separate pairs of mention-spans that appear in the OntoNotes test set into several buckets, based on the distance between the last token of the mention-span w_i and the first token of the mention-span w_j, where w_j occurs after w_i. Each bucket contains at least 50 examples of pairs of mention-spans.

4.5 Control Tasks

To ensure that our probing model is robust, we compare its performance with a control task (Hewitt and Liang, 2019). For every pair of mention-spans $(\boldsymbol{g}_1, \boldsymbol{g}_2)$, we replace one of the span representations $\boldsymbol{g}_i$ with another $\boldsymbol{g}_i'$ randomly sampled from the data set. Note that in this control task, some information of the original mention-pairs is still preserved as the other span representation in the pair is not replaced.

5 Results and Discussion

5.1 Comparison of Probing Models

Table 1 compares the performance of the probing model using span representations fine-tuned on the OntoNotes dataset against baseline span representations and a *CNN-baseline* that utilises the original pre-trained BERT encoder. The results of the control task are reported in the bottom two lines.

The probing model suggests that span representations in *BERT-coref* encode a significant amount of coreference information, as we are able to train the model to predict whether a pair of mention-spans corefer based on their span representations alone. Both *BERT-base c2f* and *BERT-large c2f* consistently score above 90% (accuracy and F1 score) on the OntoNotes test set.

We observe that both *BERT-base c2f* and *BERT-large c2f* perform better in predicting coreference arc between a pair of mention-spans compared to their respective baselines (by 2.37 points for accuracy and 2.18 F1 points on average). We find that, although training the contextual probing model to learn contextual features for coreference arc prediction helps to encode the necessary coreference information into the baseline span representations, it still cannot outperform the probing model that utilises span representations in *BERT-coref*. This is likely caused by better coreference-related features that are learned by the BERT encoder when it is fine-tuned on OntoNotes.

We also see that fine-tuning the span representations on coreference resolution task helps encode local and long-range context inside the mention-spans efficiently. This can be observed from the performance of *CNN-baseline*, where the probing model is trained using a 1D CNN layer with kernel width of 3 and 5 to allow the model to see the contribution of local and long-range dependencies, but ultimately still underperforms compared to *BERT-coref*.

Surprisingly, our baseline span representations which were constructed from only lexical priors perform better compared to the *CNN-baseline* span representations on both metrics. We attribute this to our decision of using contextual embeddings from the final layer of pre-trained BERT, as most transferable representations from contextual encoders trained with a language modelling objective tend to occur in the intermediate layers, and that the topmost layers might be overly specialised for next-word prediction (Liu et al., 2019a; Peters et al., 2018a,b; Blevins et al., 2018; Devlin et al., 2019). This might cause the CNN layer to learn suboptimal representations of the mention-spans. The probing model that we choose is also highly selective, with

	Accuracy	F1 Score
BERT-base c2f (fine-tuned)	92.93	93.02
BERT-large c2f (fine-tuned)	93.65*	93.68*
BERT-base CNN (original, K=3)	89.51	89.91
BERT-base CNN (original, K=5)	89.04	89.28
BERT-large CNN (original, K=3)	90.27	90.35
BERT-large CNN (original, K=5)	88.09	88.28
BERT-base (original) baseline	90.37	90.65
BERT-large (original) baseline	91.47	91.69
BERT-base c2f (random)	64.83	65.17
BERT-large c2f (random)	67.53	68.36

Table 1: Comparison of the probing model's performance with various mention-span representations evaluated on the OntoNotes test set. An asterisk (*) denotes the best performance on each metric. *BERT-large c2f* improves the accuracy and F1 score over the probing baseline by 3.28% and 3.03% for the base variant, while for BERT-large baseline the improvements are 2.18% and 1.99% respectively.

selectivity of 28.1 for *BERT-base c2f* and 26.1 for *BERT-large c2f*. This also means that to achieve high accuracy, the probes must rely on coreference information encoded in the span representation.

5.2 Ablations

To examine the importance of each component in *BERT-coref* span representation, we conduct an ablation study on each part of the representation and report the accuracy and the F1 score for the probing model on the test data (Table 2).[3]

The head-finding attention mechanism is crucial for coreference-arc prediction, as it contributes the highest to the final result with 0.98 and 0.95 points for accuracy and for F1 score on average, respectively. This is consistent with previous findings from Lee et al. (2017), who shows that the attention mechanism is able to learn representations important for coreference.

We also observe that span-width embeddings play an important role in determining a coreference relation, without them the performance degrades on average by 0.4 and 0.37 for accuracy and F1. Contrary to the head-finding attention and span-width embeddings, boundary representations did not contribute much to the model's performance. We hypothesise that although boundary representations may encode a large amount of information for coreference resolution, they are not significant for coreference arc prediction, as the model does not have to predict distribution over possible spans.

5.3 Encoding Long-range Coreference

We compare how our probing model performs on various separation distances between mention-spans. Figure 2 depicts F1 scores as a function of distance between pairs of mention-spans. Although performance with BERT models degrades with larger distances, the span representations in *BERT-coref* hold up better in general compared to the baseline or *CNN-baseline*. The BERT-base variant experiences a minor degradation in performance up to 5 points when $d = 125$ tokens, while for BERT-large the F1 score drops only by 7 points between $d = 0$ tokens and $d = 250$ tokens, which suggests that the depth of the Transformer layer helps to encode long-range coreference.

However, we lack sufficient evidence to suggest that the span representations are able to encode long-range coreference relations efficiently, seeing that although the encoder has been fine-tuned on OntoNotes, the model still cannot perform consistently across distant spans, with the lowest F1 score of 67% and 75% for BERT-base and BERT-large respectively, when $d = 451$ to 475 tokens.

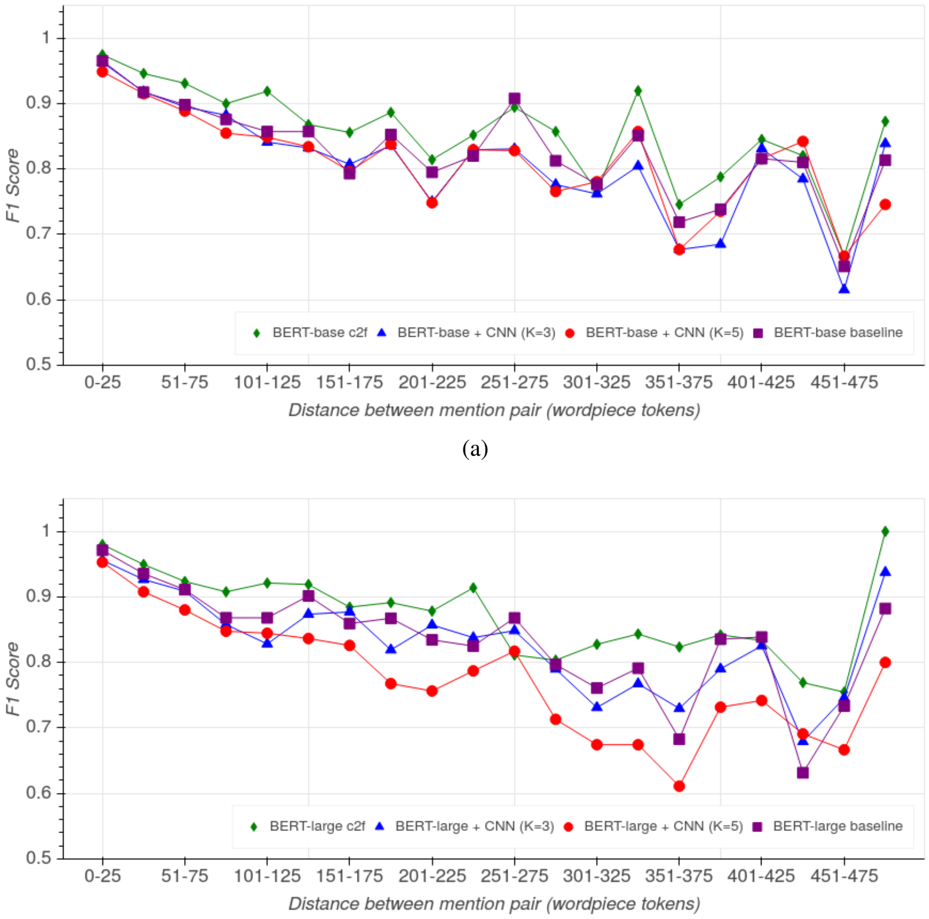
5.4 Error Analysis

We provide qualitative error analysis for predicted coreference between mention-pairs. We look at the output of *BERT-base c2f* (cased, fine-tuned) and *BERT-large c2f* (cased, fine-tuned). The predictions of both models on the same subset of 1,250 predictions from the test set are analysed. Overall, we found 93 errors in the model with BERT-base embeddings and 84 for the model with BERT-large embeddings. The errors are grouped into: *Similar Word Forms, Anaphora, Gender, Mention Para-*

[3] Results for replication experiments after acceptance are reported in Appendix A.

	Accuracy	F1 Score	ΔAccuracy	ΔF1 Score
BERT-base c2f (fine-tuned)	92.93	93.02		
- boundary representations	92.88	92.96	−0.05	−0.06
- head-finding attention	92.05	92.16	−0.88	−0.86
- span-width embeddings	92.46	92.56	−0.47	−0.46
BERT-large c2f (fine-tuned)	93.65	93.68		
- boundary representations	93.47	93.49	−0.18	−0.19
- head-finding attention	92.57	92.65	−1.08	−1.03
- span-width embeddings	93.32	93.41	−0.33	−0.27

Table 2: Comparison of the probing models on the OntoNotes test set with various components removed. The head-finding attention and span-width embeddings contribute significantly to the performance of the probing model.



(a)

(b)

Figure 2: F1 scores of the probing model as a function of separating distance between two mention-spans with BERT-base (2a) and BERT-large (2b) on test set. The performance of the model with either BERT-base or BERT-large embeddings tends to decrease as the distance between wordpiece tokens increases.

phrasing, and *Temporal and Spacial Agreement*. Although *Gender* can be considered as a subcategory of *Anaphora*, we decided to separate it to check whether gender bias is present in the models.

Table 3 portrays an overview of the errors made by both models in each category. We note that mentions separated by a distance of more than 25 tokens have a higher error rate than mentions separated by smaller distances, suggesting that *BERT-base c2f* and *BERT-large c2f* perform better when resolving coreference locally.

In the gender category, we only found one problematic example. The proper name *Scooter Libby* is consistently predicted to corefer with *she* and *her*, although the real world referent is male.

Consistent with Joshi et al. (2019), the most difficult case for both models is anaphora, even at very short distances between mentions, as in the follow-

Category	Snippet	BERT-base c2f	BERT-large c2f
Similar Word Forms	... in some of the questioning eh of *Miller*, I think ... you have *Judy Miller* there (13) ... this is the Dick Cheney aide she **agreed** to refer ... I think the **agreement** was strange (85)	17	13
Anaphora	... it was very prompt with *traffic management and emergency repair* ... ah, because *it* involved various (5) ... the news on the day of *the accident* ... instead of the east and *it* did not (277)	47	41
Gender	... killed a piece written by a reporter about **Scooter Libby** ... They didn' t say that you know until **she** walked out (58)	0	2
Mention Paraphrasing	When someone sews a patch over **a hole in an old coat**, they ... If they do, **the patch** will shrink (22) ... read a statement from **a Sixty Minutes spokesman** ... When **Mister Carson the representative** spoke ... (241)	20	19
Temporal and Spacial	... people from economic circles, who even predicted that in *1998* ... They pointed out that, *this year*, except ... (13) ... and only 582 million US dollars **last year**... momentum can not be restrained, **this year** ... (379)	9	9
Total		93	84

Table 3: Number of errors by the *BERT-base c2f* and *BERT-large c2f* fine-tuned models. The number of tokens between the highlighted mentions is given in the parenthesis. False positives are denoted **bold**, false negatives in *italic*.

ing example with a distance of only 5 tokens: “we should say it was very prompt with *traffic management and emergency repair*, ah, because *it* involved various [. . .]”. Cases of coreference between two pronouns are also difficult for both models.

The similar word forms category concerns errors in mentions with morphologically related word forms which are identified as coreferent, for instance “[. . .] this is the Dick Cheney aide she *agreed* to refer [. . .]”. I think the *agreement* was strange [. . .]”. In contrast, together with anaphora, errors involving paraphrasing and temporal and spacial agreement have an extra level of complexity in that they involve real world knowledge. For instance, for humans it is trivial that *1996* and *1997* are years and that they are different ones. The systems, on the other hand, consistently label them as coreferent, as if they were morphologically related forms.

6 Conclusion and Future Work

In this paper, we quantify the coreference information in the span representation by how well they can do on the coreference arc prediction task. We demonstrate that using mention-span representations as inputs, a simple probing model can be used to predict coreference for pairs of mention spans with accuracy and F1 score over 90%. This suggests that a significant amount of coreference information is encoded in mention-span representations obtained from BERT embeddings, which are fine-tuned on the OntoNotes dataset. Consistently with non-neural architectures, our analysis also shows that non-local coreference is challenging for span representations. Furthermore, we show that the head-finding attention mechanism encodes essential coreference-related features in span representations, even when using original pre-trained BERT embeddings.

The findings we report are solely based on an English corpus. Other pieces of research (Azerkovich, 2020; Hint et al., 2020) suggest that such positive results might be more challenging to achieve for morphologically or syntactically complex languages.

Although we work with the OntoNotes dataset, there are other challenging coreference resolution datasets that focus on ambiguous pronouns (GAP by Webster et al. (2018)) or commonsense reasoning (WinoGrande by Sakaguchi et al. (2019)), which can be used to understand coreference information in span representations better. Moreover, we would like to probe span representations derived from other pre-trained language models such as RoBERTa (Liu et al., 2019b) and SpanBERT (Joshi et al., 2020). Alternative Transformer-based architecture that is better at handling long sequences such as Longformer (Beltagy et al., 2020) also seems promising to explore, as it might improve span representations capability to model long-range coreference. Lastly, instead of building span representations from the final layer of a pre-trained BERT model, one can opt to use the activations from the intermediate layers as well as ELMo-style scalar mixing (Tenney et al., 2019; Peters et al., 2018a). We leave this to future work.

Acknowledgements

We thank the anonymous reviewers for their critical reading of our manuscript and their insightful comments and suggestions.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *CoRR*, abs/1608.04207.

Ilya Azerkovich. 2020. Using semantic information for coreference resolution with neural networks in russian. In *Analysis of Images, Social Networks and Texts*, pages 85–93, Cham. Springer International Publishing.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep RNNs encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Melbourne, Australia. Association for Computational Linguistics.

François Chollet et al. 2015. Keras. `https://keras.io`.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does bert look at? an analysis of bert's attention. In *BlackBoxNLP@ACL*.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 1026–1034, USA. IEEE Computer Society.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Helen Hint, Tiina Nahkola, and Renate Pajusalu. 2020. Pronouns as referential devices in estonian, finnish, and russian. *Journal of Pragmatics*, 155:43 – 63.

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.

Mandar Joshi, Omer Levy, Daniel S. Weld, and Luke Zettlemoyer. 2019. BERT for coreference resolution: Baselines and analysis. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA. Omnipress.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. 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.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. 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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *ArXiv*, abs/1907.10641.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguou. In *Transactions of the ACL*, page to appear.

A Appendix

A.1 Averaged Accuracy and F1 Score for Ablation Study

	Accuracy	ΔAccuracy
BERT-base c2f (fine-tuned)	92.93	
- boundary representations	92.77	−0.16
- head-finding attention	91.74	−1.19
- span-width embeddings	92.59	−0.34
BERT-large c2f (fine-tuned)	93.65	
- boundary representations	93.88	+0.23
- head-finding attention	92.65	−1.00
- span-width embeddings	93.43	−0.22

Table 4: Averaged accuracy for ablation on the OntoNotes test set. We take the average accuracy of 10 runs.

	F1 Score	ΔF1 Score
BERT-base c2f (fine-tuned)	93.02	
- boundary representations	92.89	−0.13
- head-finding attention	91.81	−1.21
- span-width embeddings	92.68	−0.34
BERT-large c2f (fine-tuned)	93.68	
- boundary representations	93.89	+0.21
- head-finding attention	92.64	−1.04
- span-width embeddings	93.44	−0.24

Table 5: Averaged F1 score for ablation on the OntoNotes test set. We take the average F1 score of 10 runs.