Coreference Resolution

Exploring Span Representations in Neural Coreference Resolution

Patrick Kahardipraja, Olena Vyshnevska

1 Introduction

Coreference resolution, the task that involves determining all referring expressions that point to the same entity in a discourse model, 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. In coreference resolution task, referring expressions (i.e. mentions) could be a common noun, a proper noun, or a pronoun, which refer to a real-world entity known as the referent. The main goal of a coreference resolution system is to group the mentions such that the corefering mentions are placed together in a coreference cluster.

In the recent years neural-based coreference resolution approaches gained prominence. This can be attributed to the ability of neural networks to learn features automatically with multiple, hierarchical representations as well as their high performance and ability to generalise better compared to rule-based approach. Most of the state-of-the-art deep learning-based coreference resolvers learn representations from word embeddings on a phrase level by constructing them explicitly (Lee, L. He, Lewis, et al. 2017) and continually refining them (Lee, L. He, and Zettlemoyer 2018; Kantor and Globerson 2019).

Although such models are powerful, especially with more advanced language representation methods (Devlin et al. 2019; Peters, Neumann, Iyyer, et al. 2018), questions still remain regarding the interpretability of neural NLP models. A wave of recent work has tried to inspect neural NLP models to understand whether they are truly able to capture linguistic information by trying to associate neural network components with distinct types of linguistic phenomena. Such line of work (Shi, Padhi, and Knight 2016; N. F. Liu et al. 2019; Tenney et al. 2019) demonstrates that deep learning-based models can encode a wide variety of syntactic and semantic information.

We follow along this line of work, focusing on the BERT model for coreference resolution (Joshi, Levy, et al. 2019) and utilise the probing task (Tenney et al. 2019; N. F. Liu et al. 2019) to find out to what degree the coreference information is encoded in the span representations first proposed by Lee, L. He, Lewis, et al. (2017). We also intend to learn how fine-tuning BERT on coreference resolution affects the linguistic knowledge learned. Additionally, we inspect if non-local coreference relations can be predicted as correctly as local ones. In order to answer these questions, we generate mention-span representations with BERT embeddings fine-tuned on OntoNotes dataset (Pradhan et al. 2012) and train a model to make predictions from the mention-span representations alone. If we could train a simple model to predict coreference relation for pairs of mentions from their span representations alone, then we could reasonably deduce that span representations encode coreference information.

Our experimental results show that a linear probing model trained on top of span representations obtained from fine-tuned BERT on coreference resolution consistently achieves > 90% accuracy and F1 on OntoNotes test set. This findings demonstrate that span representations are able to encode a significant amount of coreference information. It also suggests that fine-tuning a BERT model greatly helps with encoding coreference relations. Furthermore, by ablating components of the span representation, we found that head-finding attention mechanism plays a crucial part in encoding important coreference information. We also show that despite using fine-tuned BERT, the span representations cannot capture non-local coreference relation efficiently. Our implementation is publicly available on GitHub. [1].

[1] https://github.com/alyonavyshnevska/bert_for_coreference_resolution

2 Related Work

2.1 Approaches to Coreference Resolution

Most of the earlier systems that attempt to solve coreference resolution are rule-based. Hobbs (1978) proposed a naïve tree-search algorithm for identifying antecedents with world knowledge-based constraints to prune the antecedent search space, which started a long series of syntax-based methods. Lappin and Leass (1994) introduced an algorithm that maintains a discourse model consisting of potential antecedent references, where each antecedent is assigned a salience value based on several features. All the salience value are then summed to score and predict the most appropriate antecedent. Furthermore, Kennedy and Boguraev (1996) extended the approach to avoid full syntactic parsing by using part-of-speech tagger augmented with annotations of grammatical functions.

The field of coreference resolution underwent a shift since the first machine learning-based approach for coreference resolution was published (Connolly, Burger, and Day 1994). While rule-based systems have witnessed a short-lived revival in the 2010s (Zhou, Ticrea, and Hovy 2004; Haghighi and Klein 2009), their struggles with semantic understanding of the models led to their eventual demise. Architectures for machine learning-based coreference resolution models can be categorised as (1) mention-pair classifiers (Ng and Cardie 2002; Bengtson and Roth 2008), (2) mention-ranking models (Durrett and Klein 2013; Wiseman, Rush, S. Shieber, et al. 2015; Clark and Manning 2016a; Denis and Baldridge 2008), (3) entity-level models (Haghighi and Klein 2010; Wiseman, Rush, and S. M. Shieber 2016; Clark and Manning 2015; Clark and Manning 2016b), (4) latent-tree models (Fernandes, Santos, and Milidiu 2012; Martschat and Strube 2015; Björkelund and Kuhn 2014), and (5) span-ranking models (Lee, L. He, Lewis, et al. 2017; Lee, L. He, and Zettlemoyer 2018; Joshi, Levy, et al. 2019).

While earlier machine learning based-approaches depended on hand-crafted features, in the recent years neural-based architectures have been particularly successful in coreference resolution and NLP in general, enabled by word embeddings and contextual word representations. The first neural approach in coreference resolution was introduced by Wiseman, Rush, S. Shieber, et al. (2015), where they extended a mention-ranking model with a feed-forward neural network that can be viewed as a piecewise scoring function to determine whether a mention is anaphoric or not. Lee, L. He, Lewis, et al. (2017) proposed an end-to-end coreference resolution model that considers all spans of text, learns how to identify mentions, and clusters them together using a pairwise scoring function. Clark and Manning (2016a) proposed a novel approach to utilise reinforcement learning for coreference resolution. Using a reward-rescaled max-margin objective and REINFORCE policy gradient algorithm (Williams 1992), they exploit the independent coreference decision in mention-ranking models to directly optimise for coreference evaluation metrics. Neural coreference resolution approaches typically require different levels of semantic representations of input sentences, which is usually done by computing representations at span level given the word embeddings. We focus on these span representations and examine their capability of encoding necessary information to make coreference decisions.

2.2 Probing Tasks

Along with the rise of deep learning architectures across a plethora of NLP tasks, there are also some concerns regarding their interpretability and whether if neural-based approaches are truly able to capture linguistic concepts. 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, Padhi, and Knight (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.

N. F. Liu et al. (2019) 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 introduce coreference arc prediction task, where linear models are used to predict whether two mentions corefer. Tenney et al. (2019) introduce edge probing framework, which focuses on linguistic analysis on sub-sentence level. Their approach relies on a linear model with a projection layer and an attention mechanism on top of frozen contextual vectors to predict linguistic properties. Clark, Khandelwal, et al. (2019) further extend 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 N. F. Liu et al. (2019) and Tenney et al. (2019), but we use span representations introduced by Lee, L. He, Lewis, et al. (2017) and focus on examining coreference phenomena.

3 Overview of the Span-Ranking Architecture

The first coreference resolution model that is trained in an end-to-end manner without relying on any syntactic parsers or hand-engineered mention detectors is introduced by Lee, L. He, Lewis, et al. (2017). To construct correct coreference clusters, the neural model learns to jointly model mention detection and coreference prediction using span-ranking approach. With span-ranking approach, the model is able to consider all spans in a document up to a defined maximum length as candidate mentions and select the corresponding antecedent for each mention based on the antecedent distribution.

The span-ranking architecture estimates the joint probability distribution of mention-spans that belong to the same cluster given a document by calculating a product of multinomials for each span:

$$\begin{aligned} P(y_1, ..., y_n|D) &= \prod_{i=1}^{N} P(y_i|D) && (1)\\ &= \prod_{i=1}^{N} \frac{\text{exp}(s(i, y_i))}{\sum_{y' \in \mathcal{Y}(i)} \text{exp}(s(i, y'))} && (2) \end{aligned}$$

where $s(i, j)$ is a pairwise score for a coreference link between span i and span j in document D. The coreference score is composed of three factors: (1) $s_m(i)$, mention score of i (i.e. how possible is the span i to be a mention), (2) $s_m(j)$, mention score of j and (3) $s_a(i, j)$, joint score of i and j (i.e. assuming that i and j are both mentions, how possible that they both refer to the same entity), formulated as the following:

$$s(i, j) = \begin{cases} s_m(i) + s_m(j) + s_a(i, j) & j \neq \epsilon \\ 0 & j = \epsilon \end{cases} \tag{3}$$

where ϵ is a dummy antecedent to represent that the span is not a mention or the case where a mention is not coreferent with any of the previous mention spans.

In this model, a bidirectional LSTM (Hochreiter and Schmidhuber 1997) is used to compute vector representations $\boldsymbol{g}_i$ of each possible span i from word and character embeddings in the span. The word embeddings itself are a fixed concatenation of GloVe (Pennington, Socher, and Manning 2014) and Turian embeddings (Turian, Ratinov, and Bengio 2010). The character embeddings are learned using CNN with three various window sizes. It also provides morphological information and allows the model to deal with rare words. The span representation $\boldsymbol{g}_i$, which we will describe in detail in §4.1, is then used to compute $s_m(i)$ and $s_a(i,j)$ via feed-forward neural networks as follows:

$$s_m(i) = \boldsymbol{w}_m \cdot \text{FFNN}_m(\boldsymbol{g}_i) \tag{4}$$

$$s_a(i,j) = \boldsymbol{w}_a \cdot \text{FFNN}_a([\boldsymbol{g}_i, \boldsymbol{g}_j, \boldsymbol{g}_i \odot \boldsymbol{g}_j, \phi(i,j)]) \tag{5}$$

where $\odot$ denotes the Hadamard product between $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$, and $\phi(i,j)$ is a feature vector that encodes speaker and genre information from the metadata and the distance between a pair of spans. The final coreference score is computed through a two-stage beam search as the model complexity is quartic in terms of the document length. In the first phase, a beam of up to λT candidate mention spans is considered based on the highest mention scores $s_m(i)$, where λ is a hyperparameter and T is the document length. Afterwards, only the filtered mentions are used to compute the joint score $s_a(i,j)$ during both training and inference, with the limitation that only up to K candidate antecedents are considered for each mention.

While the first order model proposed by Lee, L. He, Lewis, et al. (2017) works quite well for coreference resolution, scoring only pairs of entity mentions means that previous coreference decisions are not taken into account when computing coreference decisions that occur afterwards. This drawback makes the model more susceptible to predicting clusters that are locally consistent but globally inconsistent. In an attempt to improve the weakness of this approach, Lee, L. He, and Zettlemoyer (2018) proposed a model that captures higher-order interactions between mention spans in predicted coreference clusters. The model utilises span-ranking architecture to refine existing span representations iteratively with the antecedent distribution as an attention mechanism. We will further refer to the model as *c2f-coref* in the paper.

In order to refine the span representations, the higher-order model estimates the antecedent distributions $P_n(y_i)$, which is obtained using the softmax function:

$$P_n(y_i) = \frac{\exp(s(\boldsymbol{g}_i^n, \boldsymbol{g}_{y_i}^n))}{\sum_{y \in \mathcal{Y}(i)} \exp(s(\boldsymbol{g}_i^n, \boldsymbol{g}_y^n))} \tag{6}$$

where s is the scoring function from Lee, L. He, Lewis, et al. (2017), $\boldsymbol{g}_i^n$ is the representation for span i at iteration n. The refined span representations also allow the model to iteratively refine $P_n(y_i)$. Afterwards, the expectation of each span i is computed by using the current antecedent distribution as the attention weight:

$$\boldsymbol{a}_i^n = \sum_{y_i \in \mathcal{Y}(i)} P_n(y_i) \cdot \boldsymbol{g}_{y_i}^n \tag{7}$$

where $\boldsymbol{a}_i^n$ can be viewed as the expected antecedent representation. The span representation $\boldsymbol{g}_i^n$ is updated with the expected antecedent representation $\boldsymbol{a}_i^n$ using coupled forget and input gates formulated as follows:

$$\boldsymbol{f}_i^n = \sigma(\mathbf{W}_f[\boldsymbol{g}_i^n, \boldsymbol{a}_i^n]) \tag{8}$$

$$\boldsymbol{g}_i^{n+1} = \boldsymbol{f}_i^n \odot \boldsymbol{g}_i^n + (\mathbf{1} - \boldsymbol{f}_i^n) \odot \boldsymbol{a}_i^n \tag{9}$$

where $\boldsymbol{f}_i^n \in [0,1]$ is a gate vector that controls whether to retain the information from the current span representation or to update the span representation with the expected antecedent. Altogether, this mechanism allows the model to compute the final antecedent distribution conditioned on up to n other mention spans.

To alleviate the additional computational cost of higher-order inference, Lee, L. He, and Zettlemoyer (2018) also proposed a coarse-to-fine beam search, which does not establish an a priori maximum coreference distance. In order to prune potential mentions aggressively, the coreference score is reformulated as follows:

$$s_c(i,j) = \boldsymbol{g}_i^\top \mathbf{W}_c \boldsymbol{g}_j \tag{10}$$

$$s(i,j) = \begin{cases} s_m(i) + s_m(j) + s_a(i,j) + s_c(i,j) & j \neq \epsilon \\ 0 & j = \epsilon \end{cases} \tag{11}$$

where $s_c(i,j)$ is a bilinear scoring function to compute a rough sketch of possible antecedents. To obtain the coreference score, the model executes a three-stage beam search. At the first stage, a beam of up to M candidate mention spans is considered based on the mention score $s_m(i)$. Afterwards, an approximation of pairwise coreference scores is computed based on $s_m(i) + s_m(j) + s_c(i,j)$ to keep top K candidate antecedents for each remaining mention span. In the last phase, the final coreference score $s(i,j)$ is computed based on the remaining pairs of mention spans.

Recently, BERT (Devlin et al. 2019) has achieved state-of-the-art results on a wide array of NLP tasks, such as question answering and natural language inference, without any substantial task-specific architecture modifications. Joshi, Levy, et al. (2019) proposed to replace the bidirectional LSTM encoder in *c2f-coref* with BERT transformers and fine-tune it on coreference resolution task. Although BERT improves the state-of-the-art results in other tasks significantly, coreference resolution still proves to be a challenging task, as BERT encoder offers only a marginal performance increase. Furthermore, the model still struggles in modelling pronouns and resolving cases where mention paraphrasing is required. We refer the model by Joshi, Levy, et al. (2019) as *BERT-coref*.

4 Probing Mention-Span Representations

4.1 Span Representations

Span representations play an important role in span-ranking models (Lee, L. He, Lewis, et al. 2017; Lee, L. He, and Zettlemoyer 2018; Joshi, Levy, et al. 2019), since they are used to compute a distribution over candidate antecedent spans. In order to be able to predict coreference relations accurately, a span representation should also capture the information of the span's internal structure and its surrounding context. For our experiments, we construct span representations which are first proposed in Lee, L. He, Lewis, et al. (2017), but with BERT embeddings (Devlin et al. 2019) instead of an LSTM-based encoder to encode lexical information of a span and its surroundings, following Joshi, Levy, 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 first and last wordpieces of the span itself. The head words are automatically learned using additive attention (Bahdanau, Cho, and Bengio 2015) over each wordpiece in a span:

$$\alpha_t = \boldsymbol{w}_\alpha \cdot \text{FFNN}_\alpha(\boldsymbol{x}_t^*) \tag{12}$$

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum\limits_{k=start(i)}^{end(i)} \exp(\alpha_k)} \tag{13}$$

$$\hat{\boldsymbol{x}}_i = \sum_{t=start(i)}^{end(i)} a_{i,t} \cdot \boldsymbol{x}_t \tag{14}$$

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:

$$\boldsymbol{g}_i = [\boldsymbol{x}^*_{start(i)}, \boldsymbol{x}^*_{end(i)}, \hat{\boldsymbol{x}}_i, \phi_i] \tag{15}$$

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.

4.2 Coreference Arc Prediction

We focus on coreference arc prediction task, which is a part of the probing tasks suite for contextual word embeddings (N. F. Liu et al. 2019; Tenney et al. 2019). In this task, a probing model is trained to determine whether two mentions refer to the same entity. As datasets for coreference resolution usually only contain annotations for corefering mentions, we produce negative samples following the approach by N. F. Liu et al. (2019). For every pair of mentions (w_i, w_j), where they belong to the same coreference cluster and w_i is an antecedent of w_j, we generate a negative example (w_{random}, w_j) where w_{random} belongs to a different coreference cluster. This method ensures that the ratio between positive and negative examples is balanced. We also follow the approach of Tenney et al. (2019) by using spans of wordpieces for mentions, as N. F. Liu et al. (2019)'s approach is limited to single-token mentions and therefore unable to fully exploit available information in a mention-span.

4.3 The Probing Model

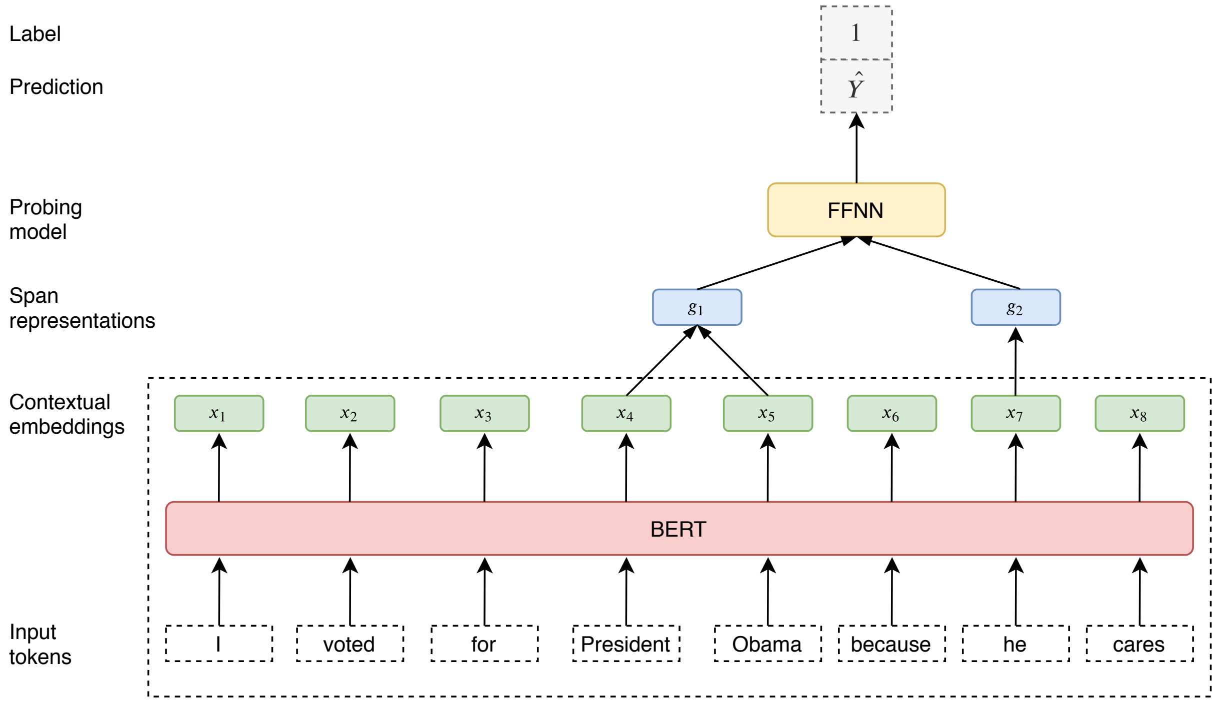


Figure 1: The probing architecture for span representations. The feed-forward neural network is trained to extract information from span representations $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$, while all the parameters inside the dashed line are frozen. The example depicts a mention-pair, where $\boldsymbol{g}_1$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

Our probing model is a simple feed-forward neural network, which is designed with a limited capacity to focus on the information that can be extracted from span representations. As the input to our probing 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 explore mention-span representations obtained from BERT. BERT (Devlin et al. 2019) is a language representation model which uses deep Transformer architecture (Vaswani et al. 2017), trained jointly with a masked language model and next sentence prediction objective on BooksCorpus (Zhu et al. 2015) and English Wikipedia, with an approximate total of 3,300M tokens. 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. We also want to compare the effect of fine-tuning BERT-based models on the quality of span representations, therefore we use two variants of BERT models: a) fine-tuned on coreference resolution task and b) without any fine-tuning.

5 Experiments

We investigate to what extent can *BERT-coref* span representations encode coreference relations. We also want to answer whether if the span representations are able to encode long-range coreference phenomena effectively, or are they simply modelling local coreference relations? Our experiments, which we describe below, are intended to analyse coreference information contained in span representations used in the span-ranking model.

5.1 Dataset

For our experiments, we use the coreference resolution annotation from the CoNLL-2012 shared task based on the OntoNotes dataset (Pradhan et al. 2012). The dataset is comprised of roughly one million words from various genres such as newswire, magazine articles, broadcast news, broadcast conversations, web data and conversational speech data, and the New Testament in English. The dataset was split into 2802 training documents, 343 validation documents, and 348 test documents. On average, the training documents contain 454 words. The largest document contains a maximum of 4009 words. Beside entities, coreference is also annotated for mentions that refer to the same event (e.g. "The European economy *grew* rapidly over the past years, *this growth* helped raising...). The main evaluation of the dataset is the average F1 score of three metrics – MUC (Vilain et al. 1995), B^3 (Bagga and Baldwin 1998) and $CEAF_{\phi}4$ (Luo 2005). Since the OntoNotes dataset only provides annotations for positive examples, we generate our own negative examples according to the aforementioned method (§4.2). We also cast the original annotations provided in the dataset into JSON format in order to work with *BERT-coref*.

5.2 Implementation 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 of the probing model are initialised with Kaiming initialisation (K. 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 model without fine-tuning the encoder weights and a BERT model that has been fine-tuned on 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, Levy, et al. (2019), which were trained using 128 wordpieces for BERT-base

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

and 384 wordpieces for BERT-large. The fine-tuned model was 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 the base and large variants as *BERT-base c2f* and *BERT-large c2f* respectively.

5.3 Baseline

As our baseline, we use span representations 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:

$$e^{(i)} = Ae^{(i)} + b \tag{16}$$

where $i = (1, 2)$, A and b are weights of the projection layer. The parameters in the projection layer are shared between representations $e^{(1)}$ and $e^{(2)}$. Afterwards, we apply self-attentional pooling operator in §4.1 over the projected representations to yield fixed-length span representations. This also 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. 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. The baseline probing architecture is depicted in Figure 2.

We compare the span representations 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 representations have a dimension of $d = 768$ for BERT-base and $d = 1024$ for BERT-large.

5.4 Long-range Coreference

We want to understand the capability of *BERT-coref* span representations. Is it able to encode long-range coreference, or is it simply modelling local coreference? We also want to find out how well the model learns from local and long-range context within the mention-spans. In order to answer both of these questions, 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 (K. He et al. 2015). Using this extended probing architecture with a CNN layer as another baseline, which we will refer to as *CNN-baseline* in this paper, enables us to examine the contribution of local and non-local context to the performance of the probing model.

To investigate whether the span representations are able to capture long-range coreference relations, we test how our probing model performs 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.

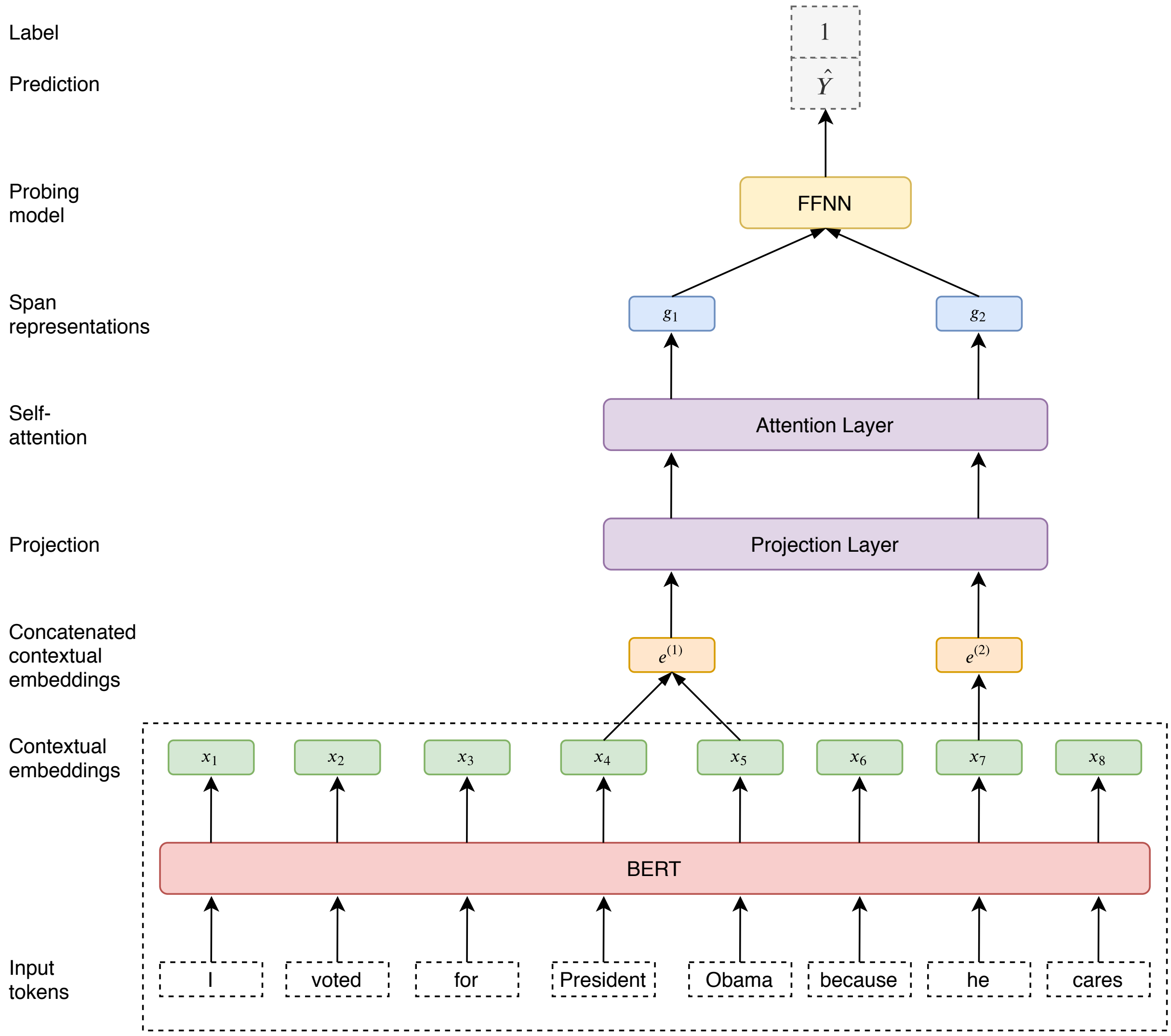


Figure 2: Probing architecture for baseline span representations. All parameters inside the dashed lines are frozen, while we train the feed-forward neural network to extract coreference information from the concatenated contextual embeddings $e^{(1)}$ and $e^{(2)}$. We used shared weights for both projection and self-attentional layer so that the model can learn the similarity between representations of mention-spans.

6 Results and Discussion

6.1 Quantitative Analysis

6.1.1 Comparison of Probing Models

We report the accuracies and the F1 scores of our probing model with span representations constructed from the BERT encoder. Table 1 compares the performance of the probing model using span representations fine-tuned on the OntoNotes dataset against baseline span representations and *CNN-baseline* that utilises the original pre-trained BERT encoder. With a simple linear probing model, we demonstrate that span representations in *BERT-coref* encode a significant amount of coreference information, as we are able to train the probing 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	F1 Score
BERT-base c2f (cased, fine-tuned)	92.93	93.02
BERT-large c2f (cased, fine-tuned)	93.65^*	93.68^*
BERT-base CNN (cased, original, kernel=3)	89.51	89.91
BERT-base CNN (cased, original, kernel=5)	89.04	89.28
BERT-large CNN (cased, original, kernel=3)	90.27	90.35
BERT-large CNN (cased, original, kernel=5)	88.09	88.28
BERT-base (cased, original) baseline	90.37	90.65
BERT-large (cased, original) baseline	91.47	91.69

Table 1: Comparison of the probing model's performance with various mention-span representations evaluated on the OntoNotes test set. 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.

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 might be caused by better coreference-related features that are learned by BERT encoder when it is fine-tuned on OntoNotes.

We find 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 performs lower 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 (N. F. Liu et al. 2019; Peters, Neumann, Iyyer, et al. 2018; Peters, Neumann, Zettlemoyer, et al. 2018; Blevins, Levy, and Zettlemoyer 2018; Devlin et al. 2019). This may cause the CNN layer to learn suboptimal representations of the mention-spans.

6.1.2 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 set of the data, which is depicted in Table 2. 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, L. He, Lewis, et al. (2017), where attention mechanism is shown to be able to learn representations that are important for coreference. We also observe that span-width embeddings play an important role in determining a coreference relation, observing that 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 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.

	Accuracy	F1 Score	ΔAccuracy	ΔF1 Score
BERT-base c2f (cased, 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 (cased, 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.

6.1.3 Encoding Long-range Coreference

We compare how our probing model performs on test set with various separation distances between mention-spans in order to learn if the span representations are able to encode long-range coreference. Figure 3 depicts F1 scores as a function of distance between pairs of mention-spans. Although performance with BERT models degrades with larger distance, 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.

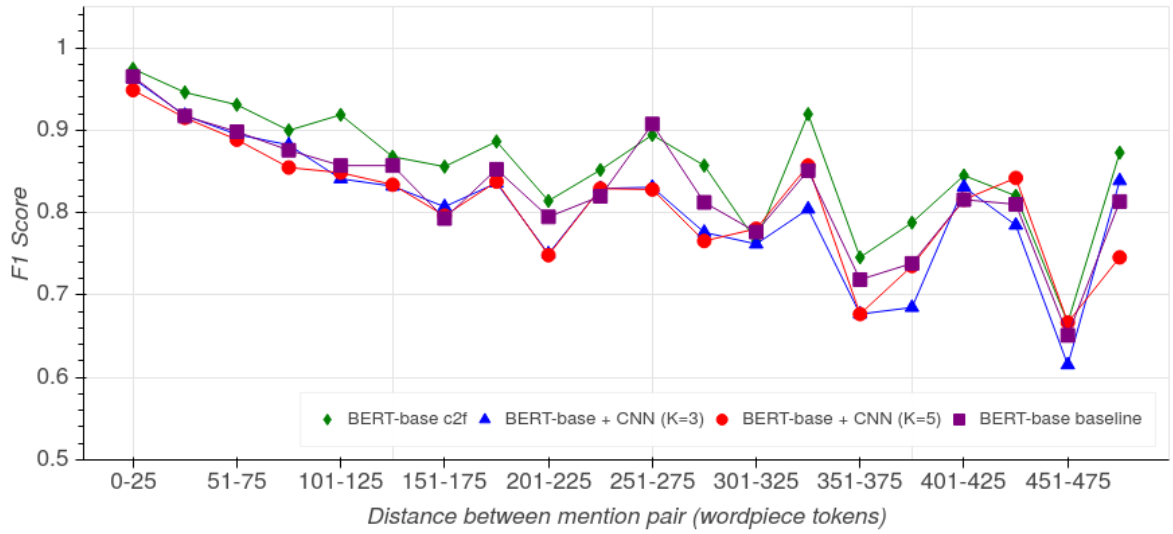
6.2 Qualitative Analysis

6.2.1 Design of Qualitative Analysis

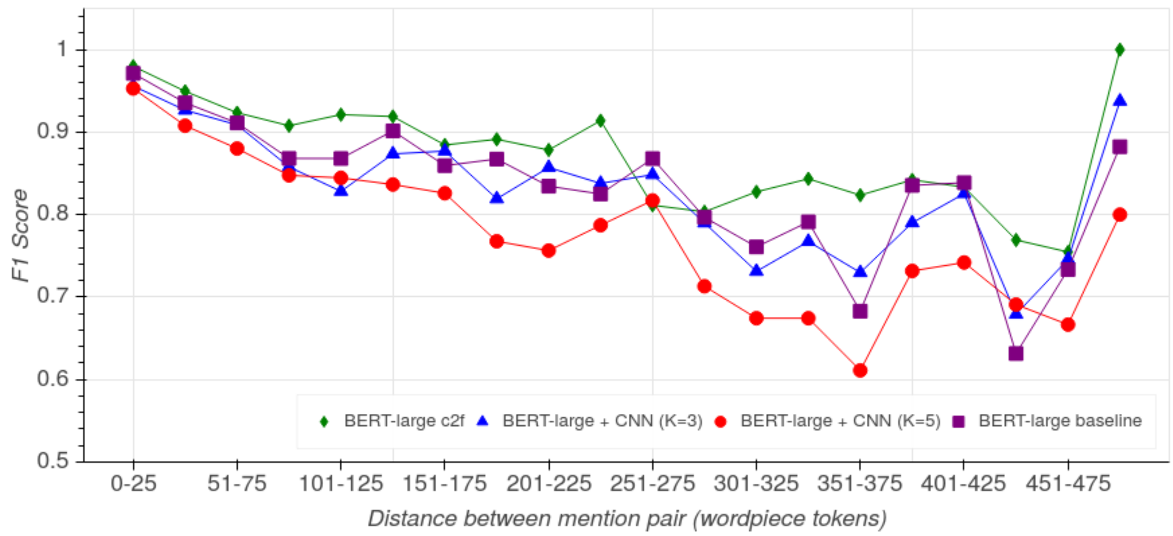
We provide qualitative error analysis for predicted coreference between mention-pairs. We identify and characterise a set of challenging errors common to the state-of-the-art systems dealing with coreference resolution in English. Such qualitative analysis enables us to detect patterns in errors that our system produces and to discuss underlying reasons for such errors.

We conduct our error analysis on the two models that are the focus of our research: *BERT-base c2f* (cased, fine-tuned) and *BERT-large c2f* (cased, fine-tuned). The predictions of both models on the same subset of 1250 predictions from the test set were 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 the following categories: *Similar Word Forms, Anaphora, Gender, Mention Paraphrasing*, and *Temporal and Spacial Agreement*. Although *Gender* can be considered as a subcategory of *Anaphora*, the decision to separate it is motivated by the curiosity to find out whether gender bias is present in the models. The raw predictions that have been analysed can be accessed from our repository.

Moreover, the errors were also categorised into *true positives, true negatives, false positives,* and *false negatives*. The numbers are provided in Table 3 and 4. We observe that false positives constitute 4.08% of the predictions for *BERT-base c2f*, while false negatives make up 3.36% of predictions of *BERT-base c2f*. *BERT-large c2f* yields only 3.2% false positives, but a slightly higher



(a)



(b)

Figure 3: F1 scores of the probing model as a function of separating distance between two mention-spans with BERT-base (3a) and BERT-large (3b) 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.

3.5% false negatives.

Tables 5, 6, and 7 portray a detailed overview of the errors made by both models in each category. As one observes immediately, mentions that are separated by a large distance have a higher error rate than mentions separated by smaller distances.

We can conclude that *BERT-base c2f* and *BERT-large c2f* perform better when resolving coreference locally. The most difficult case for all models to resolve is anaphora, which is consistent with Joshi, Levy, et al. (2019). The least problematic category for all models is gender. Out of the 1250 analysed examples only one mention led to gender-related errors – *Scooter Libby*. The mention has been consistently marked as coreferent with *she* and *her*.

	True Coreferent	True Non-Coreferent	Precision
Predicted Coreferent	575 True Positives	51 False Positives	0.92
Predicted Non-Coreferent	42 False Negatives	580 True Negatives	
Recall	0.93		

Table 3: Confusion Matrix for *BERT-base c2f*. Subset size is 1250 examples.

	True Coreferent	True Non-Coreferent	Precision
Predicted Coreferent	576 True Positives	40 False Positives	0.93
Predicted Non-Coreferent	44 False Negatives	588 True Negatives	
Recall	0.93		

Table 4: Confusion Matrix for *BERT-large c2f*. Subset size is 1250 examples.

6.2.2 Similar Word Forms

This category encloses mention pairs that share one or more lexemes. On one hand, there are examples where the tokens of one mention is a proper subset of the other mention. On the other hand, the category includes mention pairs with morphologically related word forms. False positives arise in examples such as "... this is the Dick Cheney aide she *agreed* to refer...". It was labelled as coreferent with the mention "I think the *agreement* was strange...", which appears 85 token later. False negatives often occur in examples such as "I felt it seems to be the first time that *Beijing Municipality* ... all of us are living in *Beijing* ... (32)". Here, *Beijing* is part of a larger noun phrase *Beijing Municipality* and is not recognised as coreferent with it. Similarly, in the example "... accident occurred at the main and side roads of *Jingguang Bridge, East Third Ring Road, Beijing Municipality*, resulting ... called me at noon and said something happened at *Jingguang Bridge* ... (348)", the latter mention is part of the former mention, although the distance between the two is large. As Figure 3 illustrates, the accuracy drops as the distance becomes greater.

6.2.3 Anaphora

Anaphora resolution for examples spanning over 100 tokens is intrinsically difficult, both for coreference resolution systems and for humans. For example, "... and she wouldn't share her notes with *them* although they ... She wouldn't share her notes with *her colleagues* who were writing the story (1032)". At such large distance of 1032 tokens the coreference was not recognised, even though the context of both mentions is very similar. It can be also noticed that false positives retain number and gender agreement over long-range coreference. For example, *him* and *President Clinton's* in "... *President Clinton*'s former lawyer ... former FBI director unloaded on the President who appointed *him* ... (107)" are falsely identified as coreferent, however, the two mentions do agree in number and gender. Hence, one could postulate that BERT embeddings intrinsically carry the information related to number and gender of a mention.

Anaphora resolution errors are responsible for the majority of the errors in local coreference resolution, as Table 5 illustrates. For example, *traffic management and emergency repair* is not labelled as coreferent with *it*, even though these are 5 tokens apart in "we should say it was very prompt with *traffic management and emergency repair*, ah, because *it* involved various ...". It also appears that

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) *Director Men Xianlong, Mr. Meng Xianlong, of the Command Center of the Beijing Municipal Traffic Administration* ... hear what *Director Meng said* ... (7)	3	0
Anaphora	... it was very prompt with *traffic management and emergency repair* ... ah, because *it* involved various (5) ... *the number of science and technology consulting enterprises* ... more than double of *that* ... (10)	9	5
Gender	-	0	0
Mention Paraphrasing	When someone sews a patch over **a hole in an old coat**, they ... If they do, **the patch** will shrink (22)	1	1
Temporal and Spacial	... people from economic circles, who even predicted that in *1998* ... They pointed out that, *this year*, except ... (13) ... **the Jiang-Huai area** ... experience continuous rainfall. ... **the Huang-Huai area** will bid ... (23)	1	2
Total		14	8

Table 5: Error Analysis of *BERT-base c2f* and *BERT-large c2f* models for examples with short-range coreference (0-25 tokens apart). False positives are denoted **bold**, false negatives – *italic*.

recognising coreference between two pronouns is difficult for both *BERT-base c2f* and *BERT-large c2f*. The following example demonstrates this: "... that the public has the right to know and *that* that's what *this* is all about (4)". Here, *that* and *this* have been predicted to be coreferent, however, according to the gold label they are not.

6.2.4 Gender

We made a choice to aggregate errors relating to gender pronouns in a separate category. In this subset of examples we could only observe one source of such errors – a given name *Scooter Libby* in "... one time killed a piece written by a reporter about *Scooter Libby* ... They didn't say that you know until *she* walked out (158)". It has been predicted to corefer with *she* and *her*, however the real person is a male. The name was split into four wordpiece tokens: *Sc, oot, er* and *Libby*. Interestingly, the models remained consistent in their choice of a gender pronoun for this name.

6.2.5 Mention Paraphrasing

The category consists of errors where two mentions that refer to the same entity are not recognised as coreferent as well as errors where two distinct mentions as labelled coreferent. It appears that often mentions that tend to be surrounded by similar token, such are *George Bush* and *Ronald Reagan*, are labelled coreferent as in "... If Rove gets indicted that could bring down the *Bush* administration ... the ultimate image meister for Ronald Reagan said ... (739)".

As with anaphora resolution we observe a continuous number agreement. In the example "... the peace , stability and development of the world. *Both sides* also exchanged ... six years since the establishment of diplomatic relations between *China and Uruguay* ... (164)", the two mentions are not coreferent, however, the models seem to have learned that *both sides* imply two entities. It is plausible, that in some other context *China and Uruguay* can be referred to with *both sides*.

Category	Snippet	BERT-base c2f	BERT-large c2f
Similar Word Forms	... the big news for me was Miller **recounting** what ... Tate by the way tells the Times that that **account** is false (76) ... this is the Dick Cheney aide she **agreed** to refer ... I think the **agreement** was strange (85)	3	5
Anaphora	Actually, this morning, *some listeners*, ah, happen to tell ... a lot of *them* as well as a lot of our friends ... (44) ...the reactions of citizens. – an SMS. **I** saw it ... took the subway here. Do **you** think that ... (96)	15	14
Gender	... killed a piece written by a reporter about **Scooter Libby** ... They didn' t say that you know until **she** walked out (58)	0	1
Mention Paraphrasing	... when importing and exporting **goods** ... The owners of **intellectual property rights** who ... (77) ... as many as three times the role of **Valerie Plame** ... millions of dollars in legal fees in **Miss Miller's** case. (54)	8	4
Temporal and Spacial	... the Huang-Huai area through **southern North China** ... there will also be less snow in **Northwest China** (80) ... some areas of South China, **Southwest China**, and ... southern North China and **the Huang-Huai area** ... (80)	2	2
Total		28	26

Table 6: Error Analysis of *BERT-base c2f* and *BERT-large c2f* models for examples with middle-range coreference (26-100 tokens apart). False positives are denoted **bold**, false negatives – *italic*.

Category	Snippet	BERT-base c2f	BERT-large c2f
Similar Word Forms	... they contacted **Clinton' s office back** in ... the first time that the **Clinton forces** learned about ... (370) ... of the different departments of *Beijing Municipality* ... received this order from the *municipal government*(200)	11	8
Anaphora	... took so long is that not only was **she** negotiating ... And on that point **Arianna** on that ... (180) ... the news on the day of *the accident* ... instead of the east and *it* did not (277)	23	22
Gender	... not accept the waiver of confidentiality that **Scooter Libby** offered ... her notes while they were defending **her** (1141)	0	1
Mention Paraphrasing	... but first **Judy Miller and the New York Times** finally ... in covering the controversy **the editors** killed ... (226) ... read a statement from **a Sixty Minutes spokesman** ... When **Mister Carson the representative** spoke ... (241) .	11	14
Temporal and Spacial	... and only 582 million US dollars **last year**... momentum can not be restrained, **this year** ... (379) ... News Agency, the UN, *February 13th*, Today ... multi-national forces on *the 13th* at 4 (153)	6	5
Total		51	50

Table 7: Error Analysis of *BERT-base c2f* and *BERT-large c2f* models for examples with long-range coreference (over 100 tokens apart). False positives are denoted **bold**, false negatives – *italic*.

6.2.6 Temporal and Spacial Agreement

The category aggregates errors that the models have made with regards to spacial and temporal agreement between mentions. For humans such connections are easily resolved, however the results illustrate that both *BERT-base c2f* and *BERT-large c2f* models struggle to assign a negative coreference label to, for example, *1996* and *1997*. They also fail to align *February 13th* with *the 13th* in "... Xinhua News Agency, the UN, *February 13th*, Today ... dropped by planes of the multi - national forces on *the 13th* at 4 (153)".

Moreover, *southern North China* and *Northwest China* do share two lexemes *north* and *China*, however humans would unlikely label these as corefering, whereas the models did so. Similarly, two Chinese geographical entities are predicted to be coreferent in "The east of Southwest China, as well as *the Jiang-Huai area* ... experience continuous rainfall. However, southern North China and *the Huang-Huai area* will bid ... (23)".

The examples above are clear shortcomings of the models, which might be hard to resolve through more data or more training, since even if two mentions occur in very similar contexts, they may point to entities that are usually not coreferent, such as *1996* and *1997*.

7 Conclusion and Future Work

We study coreference information in the span representations of neural-based coreference resolvers with coreference arc prediction task. We demonstrate that using mention-span representations as inputs, a simple linear probing model can be used to predict coreference for pairs of mention spans with accuracy and F1 score of 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. Our analysis also shows that span representations cannot encode non-local coreference as efficient as local coreference phenomena, based on the findings that the performance of the probing model drops to F1 score of 67% and 75% for BERT-base and BERT-large respectively when the distance between the pairs of mention spans increases to over 450 tokens. Furthermore, we discover that 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, and we acknowledge that the positive results may be a phenomenon particular to this language. Another line of work (Azerkovich 2020; Hint, Nahkola, and Pajusalu 2020) suggests that such results might be more challenging to achieve for morphologically or syntactically complex languages. One of the benefits of using neural models for coreference resolution is their ability to use contextualised word embeddings to capture similarity between words, a property that many traditional feature-based models lack. However, a similar analysis should be done for other languages before one can generalise that machine-learning based models outperform rule-based models universally.

Although we have demonstrated that the span representations incorporate a significant amount of coreference information with the OntoNotes dataset, there are other challenging coreference resolution datasets that focus on ambiguous pronouns (GAP, (Webster et al. 2018)) and commonsense reasoning (WinoGrande, (Sakaguchi et al. 2019)), which can be used to understand coreference information in span representations better. Moreover, since the publication of BERT, several other pre-trained language models that outperform BERT have been released, such as RoBERTa (Y. Liu et al. 2019) and SpanBERT (Joshi, Chen, et al. 2020). These models can be used in the place of BERT encoder to obtain better span representations. 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, Neumann, Iyyer, et al. 2018). We leave this to future work.

References

Adi, Yossi et al. (2016). "Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks". In: *CoRR* abs/1608.04207. arXiv: 1608.04207. URL: http://arxiv.org/abs/1608.04207.

Azerkovich, Ilya (2020). "Using Semantic Information for Coreference Resolution with Neural Networks in Russian". In: *Analysis of Images, Social Networks and Texts*. Ed. by Wil M. P. van der Aalst et al. Cham: Springer International Publishing, pp. 85–93. ISBN: 978-3-030-39575-9.

Bagga, Amit and Breck Baldwin (1998). "Algorithms for Scoring Coreference Chains". In: *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pp. 563–566.

Bahdanau, Dzmitry, 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*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1409.0473.

Bengtson, Eric and Dan Roth (2008). "Understanding the Value of Features for Coreference Resolution". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pp. 294–303. URL: https://www.aclweb.org/anthology/D08-1031.

Björkelund, Anders and Jonas Kuhn (June 2014). "Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 47–57. DOI: 10.3115/v1/P14-1005. URL: https://www.aclweb.org/anthology/P14-1005.

Blevins, Terra, Omer Levy, and Luke Zettlemoyer (July 2018). "Deep RNNs Encode Soft Hierarchical Syntax". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 14–19. DOI: 10.18653/v1/P18-2003. URL: https://www.aclweb.org/anthology/P18-2003.

Chollet, François et al. (2015). *Keras*. https://keras.io.

Clark, Kevin, Urvashi Khandelwal, et al. (2019). "What Does BERT Look At? An Analysis of BERT's Attention". In: *BlackBoxNLP@ACL*.

Clark, Kevin and Christopher D. Manning (July 2015). "Entity-Centric Coreference Resolution with Model Stacking". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 1405–1415. DOI: 10.3115/v1/P15-1136. URL: https://www.aclweb.org/anthology/P15-1136.

Clark, Kevin and Christopher D. Manning (2016a). "Deep Reinforcement Learning for Mention-Ranking Coreference Models". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2256–2262. DOI: 10.18653/v1/D16-1245. URL: https://www.aclweb.org/anthology/D16-1245.

Clark, Kevin and Christopher D. Manning (Aug. 2016b). "Improving Coreference Resolution by Learning Entity-Level Distributed Representations". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 643–653. DOI: 10.18653/v1/P16-1061. URL: https://www.aclweb.org/anthology/P16-1061.

Conneau, Alexis et al. (July 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)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2126–2136. DOI: 10.18653/v1/P18-1198. URL: https://www.aclweb.org/anthology/P18-1198.

Connolly, Dennis, John D. Burger, and David S. Day (1994). “A Machine Learning Approach to Anaphoric Reference”. In: *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*. ACL.

Denis, Pascal and Jason Baldridge (Oct. 2008). “Specialized Models and Ranking for Coreference Resolution”. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pp. 660–669. URL: https://www.aclweb.org/anthology/D08-1069.

Devlin, Jacob et al. (June 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Durrett, Greg and Dan Klein (2013). “Easy Victories and Uphill Battles in Coreference Resolution”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1971–1982. URL: https://www.aclweb.org/anthology/D13-1203.

Fernandes, Eraldo, Cicero dos Santos, and Roy Milidiu (2012). “Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution”. In: *Joint Conference on EMNLP and CoNLL - Shared Task*. Jeju Island, Korea: Association for Computational Linguistics, pp. 41–48. URL: https://www.aclweb.org/anthology/W12-4502.

Haghighi, Aria and Dan Klein (2009). “Simple Coreference Resolution with Rich Syntactic and Semantic Features”. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*. EMNLP 09. Singapore: Association for Computational Linguistics, pp. 1152–1161. ISBN: 9781932432633.

Haghighi, Aria and Dan Klein (2010). “Coreference Resolution in a Modular, Entity-Centered Model”. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics, pp. 385–393. URL: https://www.aclweb.org/anthology/N10-1061.

He, Kaiming et al. (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. USA: IEEE Computer Society, pp. 1026–1034. ISBN: 9781467383912. DOI: 10.1109/ICCV.2015.123. URL: https://doi.org/10.1109/ICCV.2015.123.

Hint, Helen, Tiina Nahkola, and Renate Pajusalu (2020). “Pronouns as referential devices in Estonian, Finnish, and Russian”. In: *Journal of Pragmatics* 155, pp. 43–63. ISSN: 0378-2166. DOI: https://doi.org/10.1016/j.pragma.2019.10.002.

Hobbs, Jerry R. (1978). “Resolving pronoun references”. In: *Lingua* 44.4, pp. 311–338. ISSN: 0024-3841. DOI: https://doi.org/10.1016/0024-3841(78)90006-2.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). “Long Short-Term Memory”. In: *Neural Comput.* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Joshi, Mandar, Danqi Chen, et al. (2020). “Spanbert: Improving pre-training by representing and predicting spans”. In: *Transactions of the Association for Computational Linguistics* 8, pp. 64–77.

Joshi, Mandar, Omer Levy, et al. (2019). “BERT for Coreference Resolution: Baselines and Analysis”. In: *Empirical Methods in Natural Language Processing (EMNLP)*.

Kantor, Ben and Amir Globerson (July 2019). “Coreference Resolution with Entity Equalization”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 673–677. DOI: 10.18653/v1/P19-1066. URL: https://www.aclweb.org/anthology/P19-1066.

Kennedy, C. and B.K. Boguraev (1996). “Anaphora for everyone: Pronomial anaphora resolution without a paser.” In: pp. 113–118.

Kingma, Diederik P. 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. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1412.6980.

Lappin, S. and H. Leass (1994). "An algorithm for pronominal anaphora resolution". In: vol. 20(4), pp. 535–561.

Lee, Kenton, Luheng He, Mike Lewis, et al. (Sept. 2017). "End-to-end Neural Coreference Resolution". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 188–197. DOI: 10.18653/v1/D17-1018. URL: https://www.aclweb.org/anthology/D17-1018.

Lee, Kenton, Luheng He, and Luke Zettlemoyer (June 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)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 687–692. DOI: 10.18653/v1/N18-2108. URL: https://www.aclweb.org/anthology/N18-2108.

Liu, Nelson F. et al. (June 2019). "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1073–1094. DOI: 10.18653/v1/N19-1112. URL: https://www.aclweb.org/anthology/N19-1112.

Liu, Yinhan et al. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv preprint arXiv:1907.11692*.

Luo, Xiaoqiang (Oct. 2005). "On Coreference Resolution Performance Metrics". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, pp. 25–32. URL: https://www.aclweb.org/anthology/H05-1004.

Martschat, Sebastian and Michael Strube (2015). "Latent Structures for Coreference Resolution". In: *Transactions of the Association for Computational Linguistics* 3, pp. 405–418. DOI: 10.1162/tacl_a_00147. URL: https://www.aclweb.org/anthology/Q15-1029.

Nair, Vinod 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. Haifa, Israel: Omnipress, pp. 807–814. ISBN: 9781605589077.

Ng, Vincent and Claire Cardie (2002). "Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution". In: *COLING 2002: The 19th International Conference on Computational Linguistics*. URL: https://www.aclweb.org/anthology/C02-1139.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

Peters, Matthew, Mark Neumann, Mohit Iyyer, et al. (June 2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://www.aclweb.org/anthology/N18-1202.

Peters, Matthew, Mark Neumann, Luke Zettlemoyer, et al. (Oct. 2018). "Dissecting Contextual Word Embeddings: Architecture and Representation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1499–1509. DOI: 10.18653/v1/D18-1179. URL: https://www.aclweb.org/anthology/D18-1179.

Pradhan, Sameer et al. (July 2012). "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes". In: *Joint Conference on EMNLP and CoNLL - Shared Task*. Jeju Island, Korea: Association for Computational Linguistics, pp. 1–40.

Sakaguchi, Keisuke et al. (2019). "WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale". In: *ArXiv* abs/1907.10641.

Shi, Xing, Inkit Padhi, and Kevin Knight (Nov. 2016). “Does String-Based Neural MT Learn Source Syntax?” In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1526–1534. DOI: 10.18653/v1/D16-1159. URL: https://www.aclweb.org/anthology/D16-1159.

Tenney, Ian et al. (2019). “What do you learn from context? Probing for sentence structure in contextualized word representations”. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SJzSgnRcKX.

Turian, Joseph, Lev-Arie Ratinov, and Yoshua Bengio (July 2010). “Word Representations: A Simple and General Method for Semi-Supervised Learning”. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, pp. 384–394. URL: https://www.aclweb.org/anthology/P10-1040.

Vaswani, Ashish et al. (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Vilain, Marc et al. (1995). “A Model-Theoretic Coreference Scoring Scheme”. In: *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. URL: https://www.aclweb.org/anthology/M95-1005.

Webster, Kellie et al. (2018). “Mind the GAP: A Balanced Corpus of Gendered Ambiguou”. In: *Transactions of the ACL*, to appear.

Williams, Ronald J. (May 1992). “Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning”. In: *Mach. Learn.* 8.3–4, pp. 229–256. ISSN: 0885-6125. DOI: 10.1007/BF00992696. URL: https://doi.org/10.1007/BF00992696.

Wiseman, Sam, Alexander M. Rush, and Stuart M. Shieber (2016). “Learning Global Features for Coreference Resolution”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 994–1004. DOI: 10.18653/v1/N16-1114. URL: https://www.aclweb.org/anthology/N16-1114.

Wiseman, Sam, Alexander M. Rush, Stuart Shieber, et al. (2015). “Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 1416–1426. DOI: 10.3115/v1/P15-1137. URL: https://www.aclweb.org/anthology/P15-1137.

Zhou, Liang, Miruna Ticrea, and Eduard Hovy (2004). “Multi-Document Biography Summarization”. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 434–441.

Zhu, Yukun et al. (2015). “Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books”. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. ICCV ’15. USA: IEEE Computer Society, pp. 19–27. ISBN: 9781467383912. DOI: 10.1109/ICCV.2015.11. URL: https://doi.org/10.1109/ICCV.2015.11.