BERT for Coreference Resolution: Baselines and Analysis

Mandar Joshi† Omer Levy§ Daniel S. Weld†ϵ Luke Zettlemoyer†§

† Allen School of Computer Science & Engineering, University of Washington, Seattle, WA
{mandar90,weld,lsz}@cs.washington.edu

ϵ Allen Institute of Artificial Intelligence, Seattle
{danw}@allenai.org

§ Facebook AI Research, Seattle
{omerlevy,lsz}@fb.com

Abstract

We apply BERT to coreference resolution, achieving strong improvements on the OntoNotes (+3.9 F1) and GAP (+11.5 F1) benchmarks. A qualitative analysis of model predictions indicates that, compared to ELMo and BERT-base, BERT-large is particularly better at distinguishing between related but distinct entities (e.g., President and CEO). However, there is still room for improvement in modeling document-level context, conversations, and mention paraphrasing. Our code and models are publicly available1.

1 Introduction

Recent BERT-based models have reported dramatic gains on multiple semantic benchmarks including question-answering, natural language inference, and named entity recognition (Devlin et al., 2019). Apart from better bidirectional reasoning, one of BERT's major improvements over previous methods (Peters et al., 2018; McCann et al., 2017) is passage-level training,2 which allows it to better model longer sequences.

We fine-tune BERT to coreference resolution, achieving strong improvements on the GAP (Webster et al., 2018) and OntoNotes (Pradhan et al., 2012) benchmarks. We present two ways of extending the c2f-coref model in Lee et al. (2018). The independent variant uses non-overlapping segments each of which acts as an independent instance for BERT. The overlap variant splits the document into overlapping segments so as to provide the model with context beyond 512 tokens. BERT-large improves over ELMo-based c2f-coref 3.9% on OntoNotes and 11.5% on GAP (both absolute).

A qualitative analysis of BERT and ELMo-based models (Table 3) suggests that BERT-large (unlike BERT-base) is remarkably better at distinguishing between related yet distinct entities or concepts (e.g., Repulse Bay and Victoria Harbor). However, both models often struggle to resolve coreferences for cases that require world knowledge (e.g., the developing story and the scandal). Likewise, modeling pronouns remains difficult, especially in conversations.

We also find that BERT-large benefits from using longer context windows (384 word pieces) while BERT-base performs better with shorter contexts (128 word pieces). Yet, both variants perform much worse with longer context windows (512 tokens) in spite of being trained on 512-size contexts. Moreover, the overlap variant, which artificially extends the context window beyond 512 tokens provides no improvement. This indicates that using larger context windows for pretraining might not translate into effective long-range features for a downstream task. Larger models also exacerbate the memory-intensive nature of span representations (Lee et al., 2017), which have driven recent improvements in coreference resolution. Together, these observations suggest that there is still room for improvement in modeling document-level context, conversations, and mention paraphrasing.

2 Method

For our experiments, we use the higher-order coreference model in Lee et al. (2018) which is the current state of the art for the English OntoNotes dataset (Pradhan et al., 2012). We refer to this as c2f-coref in the paper.

2.1 Overview of c2f-coref

For each mention span x, the model learns a distribution $P(\cdot)$ over possible antecedent spans $Y$:  

---

1https://github.com/mandarjoshi90/coref

2Each BERT training example consists of around 512 word pieces, while ELMo is trained on single sentences.
The scoring function $s(x,y)$ between spans $x$ and $y$ uses fixed-length span representations, $g_x$ and $g_y$, to represent its inputs. These consist of a concatenation of three vectors: the two LSTM states of the span endpoints and an attention vector computed over the span tokens. It computes the score $s(x,y)$ by the mention score of $x$ (i.e. how likely is the span $x$ to be a mention), the mention score of $y$, and the joint compatibility score of $x$ and $y$ (i.e. assuming they are both mentions, how likely are $x$ and $y$ to refer to the same entity). The components are computed as follows:

$$s(x,y) = s_m(x) + s_m(y) + s_c(x,y)$$

$$s_m(x) = \text{FFNN}_m(g_x)$$

$$s_c(x,y) = \text{FFNN}_c(g_x, g_y, \phi(x,y))$$

where FFNN$(\cdot)$ represents a feedforward neural network and $\phi(x,y)$ represents speaker and metadata features. These span representations are later refined using antecedent distribution from a span-ranking architecture as an attention mechanism.

### 2.2 Applying BERT

We replace the entire LSTM-based encoder (with ELMo and GloVe embeddings as input) in c2f-coref with the BERT transformer. We treat the first and last word-pieces (concatenated with the attended version of all word pieces in the span) as span representations. Documents are split into segments of $\text{max}_\text{segment}_\text{len}$, which we treat as a hyperparameter. We experiment with two variants of splitting:

**Independent** The independent variant uses non-overlapping segments each of which acts as an independent instance for BERT. The representation for each token is limited to the set of words that lie in its segment. As BERT is trained on sequences of at most 512 word pieces, this variant has limited encoding capacity especially for tokens that lie at the start or end of their segments.

**Overlap** The overlap variant splits the document into overlapping segments by creating a $T$-sized segment after every $T/2$ tokens. These segments are then passed on to the BERT encoder independently, and the final token representation is derived by element-wise interpolation of representations from both overlapping segments.

Let $r_1 \in \mathbb{R}^d$ and $r_2 \in \mathbb{R}^d$ be the token representations from the overlapping BERT segments. The final representation $r \in \mathbb{R}^d$ is given by:

$$f = \sigma(w^T[r_1; r_2])$$

$$r = f \cdot r_1 + (1 - f) \cdot r_2$$

where $w \in \mathbb{R}^{2d \times d}$ is a trained parameter and $[;]$ represents concatenation. This variant allows the model to artificially increase the context window beyond the $\text{max}_\text{segment}_\text{len}$ hyperparameter.

All layers in both model variants are then fine-tuned following Devlin et al. (2019).

### 3 Experiments

We evaluate our BERT-based models on two benchmarks: the paragraph-level GAP dataset (Webster et al., 2018), and the document-level English OntoNotes 5.0 dataset (Pradhan et al., 2012). OntoNotes examples are considerably longer and typically require multiple segments to read the entire document.

#### Implementation and Hyperparameters

We extend the original Tensorflow implementations of c2f-coref$^3$ and BERT.$^4$ We fine tune all models on the OntoNotes English data for 20 epochs using a dropout of 0.3, and learning rates of $1 \times 10^{-5}$ and $2 \times 10^{-4}$ with linear decay for the BERT parameters and the task parameters respectively. We found that this made a sizable impact of 2-3% over using the same learning rate for all parameters.

We trained separate models with $\text{max}_\text{segment}_\text{len}$ of 128, 256, 384, and 512; the models trained on 128 and 384 word pieces performed the best for BERT-base and BERT-large respectively. As span representations are memory intensive, we truncate documents randomly to eleven segments for BERT-base and 3 for BERT-large during training. Likewise, we use a batch size of 1 document following (Lee et al., 2018). While training the large model requires 32GB GPUs, all models can be tested on 16GB GPUs. We use the cased English variants in all our experiments.

#### Baselines

We compare the c2f-coref + BERT system with two main baselines: (1) the original ELMo-based c2f-coref system (Lee et al., 2018), and (2) its predecessor, e2e-coref (Lee et al.,

---

$^3$https://github.com/kentonl/e2e-coref/

$^4$https://github.com/google-research/bert
Table 1: OntoNotes: BERT improves the c2f-coref model on English by 0.9% and 3.9% respectively for base and large variants. The main evaluation is the average F1 of three metrics – MUC, B₃, and CEAFφ₄ on the test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>M</th>
<th>F</th>
<th>B</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>e2e-coref</td>
<td>67.2</td>
<td>62.2</td>
<td>0.92</td>
<td>64.7</td>
</tr>
<tr>
<td>c2f-coref</td>
<td>75.8</td>
<td>71.1</td>
<td>0.94</td>
<td>73.5</td>
</tr>
<tr>
<td>BERT + RR Liu et al. (2019)</td>
<td>80.3</td>
<td>77.4</td>
<td>0.96</td>
<td>78.8</td>
</tr>
<tr>
<td>BERT-base + c2f-coref</td>
<td>84.4</td>
<td>81.2</td>
<td>0.96</td>
<td>82.8</td>
</tr>
<tr>
<td>BERT-large + c2f-coref</td>
<td>86.9</td>
<td>83.0</td>
<td>0.95</td>
<td>85.0</td>
</tr>
</tbody>
</table>

Table 2: GAP: BERT improves the c2f-coref model by 11.5%. The metrics are F1 score on Masculine and Feminine examples, Overall, and a Bias factor (F / M).

3.2 Document Level: OntoNotes

OntoNotes (English) is a document-level dataset from the CoNLL-2012 shared task on coreference resolution. It consists of about one million words of newswire, magazine articles, broadcast news, broadcast conversations, web data and conversational speech data, and the New Testament. The main evaluation is the average F1 of three metrics – MUC, B₃, and CEAFφ₄ on the test set according to the official CoNLL-2012 evaluation scripts.

Table 1 shows that BERT-base offers an improvement of 0.9% over the ELMo-based c2f-coref model. Given how gains on coreference resolution have been hard to come by as evidenced by the table, this is still a considerable improvement. However, the magnitude of gains is relatively modest considering BERT’s arguably better architecture and many more trainable parameters. This is in sharp contrast to how even the base variant of BERT has very substantially improved the state of the art in other tasks. BERT-large, however, improves c2f-coref by the much larger margin of 3.9%. We also observe that the overlap variant offers no improvement over independent.

Concurrent with our work, Kantor and Globerson (2019), who use higher-order entity-level representations over “frozen” BERT features, also report large gains over c2f-coref. While their feature-based approach is more memory efficient, the fine-tuned model seems to yield better results. Also concurrent, SpanBERT (Joshi et al., 2019), another self-supervised method, pretrains span representations achieving state of the art results (Avg. F1 79.6) with the independent variant.
### 4 Analysis

We performed a qualitative comparison of ELMo and BERT models (Table 3) on the OntoNotes English development set by manually assigning error categories (e.g., pronouns, mention paraphrasing) to incorrect predicted clusters. Overall, we found 93 errors for BERT-base and 74 for BERT-large from the same 15 documents.

#### Strengths

We did not find salient qualitative differences between ELMo and BERT-base models, which is consistent with the quantitative results (Table 1). BERT-large improves over BERT-base in a variety of ways including pronoun resolution and lexical matching (e.g., *race track* and *track*).

In particular, the BERT-large variant is better at distinguishing related, but distinct, entities. Table 3 shows several examples where the BERT-base variant merges distinct entities (like *Ocean Theater* and *Marine Life Center*) into a single cluster. BERT-large seems to be able to avoid such merging on a more regular basis.

#### Weaknesses

An analysis of errors on the OntoNotes English development set suggests that better modeling of document-level context, conversations, and entity paraphrasing might further improve the state of the art.

*Longer documents* in OntoNotes generally contain larger and more spread-out clusters. We focus on three observations – (a) Table 4 shows how models perform distinctly worse on longer documents, (b) both models are unable to use larger segments more effectively (Table 5) and perform worse when the *max_segment_len* of 512 is used, and, (c) using overlapping segments to provide additional context does not improve results (Table 1). Recent work (Joshi et al., 2019) suggests that BERT’s inability to use longer sequences effectively is likely a by-product of training on short sequences.

Comparing preferred segment lengths for base and large variants of BERT indicates that larger

<table>
<thead>
<tr>
<th>Category</th>
<th>Snippet</th>
<th>#base</th>
<th>#large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Related Entities</td>
<td>Watch spectacular performances by dolphins and sea lions at the <em>Ocean Theater</em>...</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Lexical</td>
<td>Over the past 28 years, <em>the Ocean Park</em> has basically... <em>The entire park</em> has been...</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>Pronouns</td>
<td>In the meantime, our children need <em>an education</em>. <em>That</em>’s all we’re asking.</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>Mention Paraphrasing</td>
<td>Today Britain’s <em>Prince Charles and his wife Camilla</em>...</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>Conversation</td>
<td><em>(Priscilla:)</em> My mother was Thelma Wahl. She was ninety years old... <em>(Keith:)</em> Priscilla Scott is mourning. Her mother Thelma Wahl was a resident...</td>
<td>18</td>
<td>16</td>
</tr>
<tr>
<td>Misc.</td>
<td>He is <em>my</em> Goddess, ah</td>
<td>17</td>
<td>17</td>
</tr>
</tbody>
</table>

**Table 3: Qualitative Analysis** #base and #large refers to the number of cluster-level errors on a subset of the OntoNotes English development set. Underlined and **bold-faced** mentions respectively indicate incorrect and missing assignments to *italicized* mentions/clusters. The miscellaneous category refers to other errors including (reasonable) predictions that are either missing from the gold data or violate annotation guidelines.

<table>
<thead>
<tr>
<th>Doc length</th>
<th>#Docs</th>
<th>Spread</th>
<th>F1 (base)</th>
<th>F1 (large)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 128</td>
<td>48</td>
<td>37.3</td>
<td>80.6</td>
<td>84.5</td>
</tr>
<tr>
<td>128 - 256</td>
<td>54</td>
<td>71.7</td>
<td>80.0</td>
<td>83.0</td>
</tr>
<tr>
<td>256 - 512</td>
<td>74</td>
<td>109.9</td>
<td>78.2</td>
<td>80.0</td>
</tr>
<tr>
<td>512 - 768</td>
<td>64</td>
<td>155.3</td>
<td>76.8</td>
<td>80.2</td>
</tr>
<tr>
<td>768 - 1152</td>
<td>61</td>
<td>197.6</td>
<td>71.1</td>
<td>76.2</td>
</tr>
<tr>
<td>1152+</td>
<td>42</td>
<td>255.9</td>
<td>69.9</td>
<td>72.8</td>
</tr>
<tr>
<td>All</td>
<td>343</td>
<td>179.1</td>
<td>74.3</td>
<td>77.3</td>
</tr>
</tbody>
</table>

**Table 4: Performance** on the English OntoNotes dev set generally drops as the document length increases. Spread is measured as the average number of tokens between the first and last mentions in a cluster.

<table>
<thead>
<tr>
<th>Segment Length</th>
<th>F1 (BERT-base)</th>
<th>F1 (BERT-large)</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>74.4</td>
<td>76.6</td>
</tr>
<tr>
<td>256</td>
<td>73.9</td>
<td>76.9</td>
</tr>
<tr>
<td>384</td>
<td>73.4</td>
<td>77.3</td>
</tr>
<tr>
<td>450</td>
<td>72.2</td>
<td>75.3</td>
</tr>
<tr>
<td>512</td>
<td>70.7</td>
<td>73.6</td>
</tr>
</tbody>
</table>

**Table 5: Performance** on the English OntoNotes dev set with varying values for *max_segment_len*. Neither model is able to effectively exploit larger segments; they perform especially badly when *maximum_segment_len* of 512 is used.

---

Each incorrect cluster can belong to multiple categories.
models might better encode longer contexts. However, larger models also exacerbate the memory-intensive nature of span representations,\footnote{We required a 32GB GPU to finetune BERT-large.} which have driven recent improvements in coreference resolution. These observations suggest that future research in pretraining methods should look at more effectively encoding document-level context using sparse representations \cite{transferring}. Modeling pronouns especially in the context of conversations (Table 3), continues to be difficult for all models, perhaps partly because c2f-coref does very little to model dialog structure of the document. Lastly, a considerable number of errors suggest that models are still unable to resolve cases requiring mention paraphrasing. For example, bridging the Royals with Prince Charles and his wife Camilla likely requires pretraining models to encode relations between entities, especially considering that such learning signal is rather sparse in the training set.

5 Related Work

Scoring span or mention pairs has perhaps been one of the most dominant paradigms in coreference resolution. The base coreference model used in this paper from \cite{Leeetal2018} belongs to this family of models \cite{NgCardie2002,BengtsonRoth2008,DenisBaldridge2008,Fernandesetal2012,DurrettKlein2013,Wisemanetal2015,ClarkManning2016,Leeetal2017}.

More recently, advances in coreference resolution and other NLP tasks have been driven by unsupervised contextualized representations \cite{Petersetal2018,Devlinetal2019,McCannetal2017,Joshietal2019}. Of these, BERT \cite{Devlinetal2019} notably uses pretraining on passage-level sequences (in conjunction with a bidirectional masked language modeling objective) to more effectively model long-range dependencies. SpanBERT \cite{Joshietal2019} focuses on pretraining span representations achieving current state of the art results on OntoNotes with the independent variant.

References


