Using Type Information to Improve Entity Coreference Resolution

Sopan Khosla   **Carolyn Rose**

Language Technologies Institute
Carnegie Mellon University, USA

{sopank, cprose}@cs.cmu.edu

Abstract

Coreference resolution (CR) is an essential part of discourse analysis. Most recently, neural approaches have been proposed to improve over SOTA models from earlier paradigms. So far none of the published neural models leverage external semantic knowledge such as type information. This paper offers the first such model and evaluation, demonstrating modest gains in accuracy by introducing either gold standard or predicted types. In the proposed approach, type information serves both to (1) improve mention representation and (2) create a soft type consistency check between coreference candidate mentions. Our evaluation covers two different grain sizes of types over four different benchmark corpora.

1 Introduction

Coreference resolution (CR) is an extensively studied problem in computational linguistics and NLP (Hobbs, 1978; Lappin and Leass, 1994; Mitkov, 1999; Ng, 2017; Clark and Manning, 2016; Lee et al., 2017). Solutions to this problem allow us to make meaningful links between concepts and entities within a discourse and therefore serves as a valuable pre-processing step for downstream tasks like summarization and question-answering (Steinberger et al., 2007; Dasigi et al., 2019; Sukthanker et al., 2020a).

Recently, multiple datasets including Ontonotes (Pradhan et al., 2012), Litbank (Bamman et al., 2020), EmailCoref (Dakle et al., 2020), and WikiCoref (Ghaddar and Langlais, 2016) have been proposed as benchmark datasets for CR, especially in the sub-area of entity anaphora (Sukthanker et al., 2020b). Entity anaphora is a simpler starting place for work on anaphora because unlike abstract anaphora (Webber, 1991), entity anaphora are pronouns or noun phrases that refer to an explicitly mentioned entity in the discourse rather than an abstract idea that must be constructed from a repackaging of information revealed over an extended text. An affordance of entity anaphora is that they have easily articulated semantic types. Most of the entity CR datasets are extensively annotated for syntactic features (like constituency parse etc.) and semantic features (like entity-types). However, none of the published SOTA methods (Lee et al., 2017; Joshi et al., 2019, 2020) explicitly leverage the type information.

In this paper, we present a proof of concept to portray the benefits of using type information in neural approaches for CR. Named entities are generally divided generically (e.g. person, organization etc.) or in a domain-specific manner (e.g. symptom, drug, test etc.). In this work, we consider CR datasets that contain generic entity-types. One challenge is that the different corpora do not utilize the same set of type tags. For example, OntoNotes includes 18 types while EmailCoref includes only 4. Thus, we evaluate the performance of the proposed modeling approach on each dataset both with the set of type tags germane to the dataset as well as a common set of four basic types (person, org, location, facility) inspired from research on Named Entity Recognition (NER) (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003).

Our motivation is similar to (Durrett and Klein, 2014), which used a structured CRF with hand-curated features to jointly-model the tasks of CR, entity typing, and entity linking. Their joint architecture showed an improved performance on CR over the independent baseline. However, our work differs from there's as we show the benefits of entity-type information in neural models that use contextualized representations like BERT (Peters et al., 2018). Some prior art (Petroni et al., 2019; Roberts et al., 2020) argues that contextual-
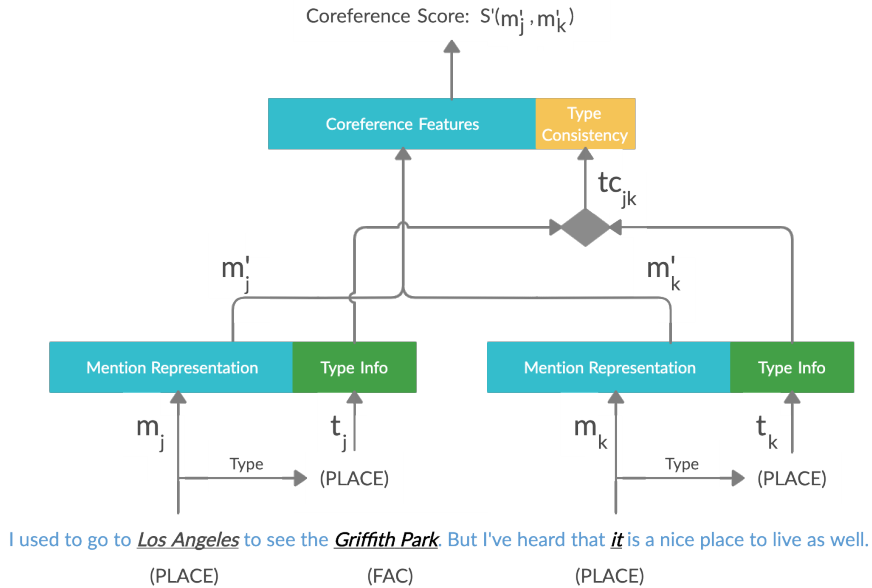
20

Figure 1: We improve Bamman et al. (2020) for entity coreference resolution by incorporating type information at two levels. (1) Type information is concatenated with the mention span representation created by their model; and (2) A consistency check is incorporated that compares the types of two mentions under consideration to calculate the coreference score. Please refer to Section 3 for details.

ized embeddings implicitly capture facts and relationships between real-world entities. However, in this work, we empirically show that access to explicit knowledge about entity-types benefits neural models that use BERT for CR. We show a consistent improvement in performance on four different coreference datasets from varied domains.

Our contribution is that we evaluate the impact of the introduction of type information in neural entity coreference at two different levels of granularity (which we refer to as original vs common), demonstrating their utility both in the case where gold standard type information is available, and the more typical case where it is predicted.

2 Related Work

Neural Coreference Resolution: Recently, neural approaches to coreference (Joshi et al., 2020, 2019; Lee et al., 2018, 2017) have begun to show their prowess. The SOTA models show impressive performance on state-of-the-art datasets like OntoNotes (Pradhan et al., 2012) and GAP (Webster et al., 2018). The notable architecture proposed by Lee et al. (2017) scores pairs of entity mentions independently and later uses a clustering algorithm to find coreference clusters. On the other hand, Lee et al. (2018) improve upon this foundation by introducing an approximated higher-order inference that iteratively updates the existing span representation using its antecedent distribution. Moreover, they propose a coarse-to-fine grained approach to pairwise scoring for tackling the computational challenges caused due to the iterative higher-order inference. More recently, Joshi et al. (2019, 2020) showed that use of contextual representations instead of word-embeddings like GloVe (Pennington et al., 2014) can further boost the results over and above those just mentioned. Our work offers additional improvement by building on the model proposed in Bamman et al. (2020), which is based on Lee et al. (2017), and adds additional nuanced information grounded in semantic types.

Type Information: Named Entity Recognition datasets (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003; Li et al., 2016) often group entity mentions into different types (or categories) depending on the domain and the potential downstream applications of the corpus. For example, the medical corpus used in the i2b2 Challenge 2010 (Uzuner et al., 2011) annotates domain-specific types like *problem*, *test*, *symptom* etc., whereas, a more general-domain dataset like CoNLL-2002 (Tjong Kim Sang, 2002) uses generic types like *person*, *organization*, and *location*. Type information as a predictive signal has been shown to be beneficial for NLP tasks like relation extraction (Soares et al., 2019) and entity-linking (Chen et al., 2020). It affords some level

of disambiguation, which assists models with filtering out some incorrect predictions in order to increase the probability of a correct prediction. In this work, we evaluate the benefits of using explicit type information for CR. We show that a model that leverages entity types associated with the anaphoric/ antecedent mentions significantly reduces the problem of type inconsistency in the output coreference clusters and thus improves the overall performance of the neural baseline on four datasets.

Type Information for CR: Multiple prior works have shown type-information to be a useful feature for shallow coreference resolution classifiers (Soon et al., 2001; Bengtson and Roth, 2008; Ponzetto and Strube, 2006; Haghighi and Klein, 2010; Durrett and Klein, 2014). (Soon et al., 2001) take the most frequent sense for each noun in WordNet as the semantic class for that noun and use a decision-tree for pairwise classification of whether two samples co-refer each other. (Bengtson and Roth, 2008) use a hypernym tree to extract the type information for different common nouns, and compare the proper names against a predefined list to determine if the mention is a person. They, then, pass this and many other features (like distance, agreement, etc.) through a regularized average perceptron for pairwise classification. This paper expands on these studies to show that entity-type information is also beneficial for neural models that use contextualized representations like BERT (Peters et al., 2018), which have been argued to implicitly capture facts and relationships between real-world entities (Petroni et al., 2019; Roberts et al., 2020).

3 Model

In this section, we explain how we introduce type information into a neural CR system.

3.1 Baseline

We use the model proposed in Bamman et al. (2020) as our baseline. The model gives state-of-the-art scores on the LitBank corpus (Bamman et al., 2020) and is an end-to-end mention ranking system based on Lee et al. (2017), which has shown competitive performance on the OntoNotes dataset. However, this model differs from Lee et al. (2017) as it uses BERT embeddings, omits author and genre information, and only focuses on the task of mention-linking. Since our main goal is to evaluate the benefits of type information, we too separate mention-linking from mention-identification and only show results computed over gold-standard mentions. This controls for the effects of the mention-identification module's performance on our experiments. Impact of type-information incorporation in the real-world end-to-end CR setting (mention identification + linking) is left as future work.

The BERT embeddings for each token i are passed through a bi-directional LSTM (x_i). To represent a mention m with start and end positions s, e respectively, x_s, x_e, attention over $x_s, ..., x_e$, and features to represent the width (wi) and inclusion within quotations (qu) are concatenated.

$$m = [x_s; x_e; Att(x_s, .., x_e); wi; qu] \quad (1)$$

Finally, given the representation of two mentions m_j and m_k, their coreference score $\mathbf{S}(m_j, m_k)$ is computed by concatenating $m_j, m_k, m_j \odot m_k$, distance (d) between the mentions and whether one mention is nested (n) within the other, which are then passed through fully-connected layers (**FC**).

$$\mathbf{S}(m_j, m_k) = \mathbf{FC}([m_j; m_k; m_j \odot m_k; d; n]) \quad (2)$$

We refer the reader to (Bamman et al., 2020; Lee et al., 2017) for more details about the architecture.

3.2 Entity Type Information

We improve the above model by including entity-type information on two levels (Figure 1). First, we concatenate the entity-type t of the mention to m (in Eq. 1) to improve the mention representation.

$$m' = [m; t] \quad (3)$$

This allows the model access to the entity type of the mention as an additional feature. We call this **+ET-self**. Second, to check the type consistency (softly) between any two mentions under consideration as possibly coreferent, we append a feature (tc) in Eq. 2, which takes the value 0 if both mentions have the same type, and 1 otherwise. For example, in Figure 1, since *Los Angeles* and *it* have the same entity-type PLACE, $tc_{jk} = 0$.

$$\mathbf{S}'(m'_j, m'_k) = \mathbf{FC}([m'_j; m'_k; m'_j \odot m'_k; d; n; tc_{jk}]) \quad (4)$$

This part of the approach is referred to as **+ET-cross** throughout the remainder of the paper. We

decide against the use of a hard consistency check (which would filter out mentions which do not have the same type) as it might not generalize well to bridging anaphora (Clark, 1975) where the anaphor refers to an object that is associated with, but not identical to, the antecedent (Poesio et al., 2018). In such cases, the type of the anaphora and its antecedent may not match. Finally, our architecture combines both components together as **+ET** (ET = ET-self + ET-cross).

4 Datasets

We gauge the benefits of using entity-type information on the four datasets discussed below.

LitBank. This dataset (Bamman et al., 2020) contains coreference annotations for 100 literary texts.[1] This dataset limits the markable mentions to six entity-types, where majority of the mentions (83.1%) point to a person.

EmailCoref. This dataset (Dakle et al., 2020) comprises of 46 email threads with a total of 245 email messages.[2] Similar to LitBank, it considers a mention to be a span of text that refers to a real-world entity. In this work, we filter out pronouns that point towards multiple entities in the email (e.g. we, they) thus only focusing on singular mentions.

Ontonotes. From this multi-lingual dataset, we evaluate on the subset (english) from OntoNotes that was used in the CoNLL-2012 shared task (Pradhan et al., 2012).[3] It contains 2802 training, 343 development, and 348 test documents. The dataset differs from LitBank in its annotation scheme with the biggest difference being the fact that it does not annotate singletons.

It contains annotations for 18 different entity-types. However, unlike LitBank and EmailCoref, not all mentions have an associated entity-type. For example, none of the pronoun mentions are given a type even if they act as anaphors to typed entities. We partially ameliorate this issue by extracting gold coreference clusters that contain at least one typed mention and assigning the majority type in that cluster to all of its elements. For example, in Figure 1, if *Los Angeles* is typed PLACE, and *it* is in the gold coreference cluster of *Los Angeles* (no other element in the cluster), then *it* is also assigned the type PLACE.

Dataset	#Types	Categories
LitBank	6	PER, LOC, FAC, GPE, VEH, ORG
EmailCoref	4	PER, ORG, LOC, DIG
WikiCoref	8	ORG, PER, CORP, EVENT, PLACE, THING, OTHER, NA
OntoNotes	19	ORG, WOA, LOC, CARD, EVENT, NORP, GPE, DATE, PER, FAC, QUANT, ORD, TIME, PROD, PERC, MON, LAW, LANG, NA
Common	5	PER, ORG, LOC, FAC, OTHER

Table 1: Type statistics for corpora used in this study.

WikiCoref. This corpus, released by (Ghaddar and Langlais, 2016), comprises 30 documents from wikipedia annotated for coreference resolution.[4] The annotations contain additional metadata, like the associated freebase rdf link for each mention (if available). We use this rdf entry to extract the mention's entity types from freebase dump. Mentions that do not get any type are marked *NA*. The first 24 documents are chosen for training, the next 3 for development, and the rest for testing.

The above-discussed datasets differ in the number as well as the categories of entity-types they originally annotate (Table 1). Apart from a common list of types (like PER, ORG, LOC), they also include corpus-specific categories like DIGital (EmailCoref), MONey, and LANG (OntoNotes). We carry out experiments with two sets of types – original and common – for each dataset. The common set of types include the following 5 categories: PER, ORG, LOC, FAC, OTHER.

5 Experiments and Results

In this section, we provide the results of our empirical experiments.

Evaluation Metrics: We convert all three datasets into the CoNLL 2012 format and report the F1 score for MUC, B^3, and CEAF metrics using the CoNLL-2012 official scripts. The performances are compared on the average F1 of the above-mentioned metrics.

For EmailCoref, OntoNotes, and WikiCoref, we report the mean score of 5 independent runs of the model with different seeds. Whereas, for LitBank, we present the 10-fold cross-validation results.[5]

---

[1]https://github.com/dbamman/lrec2020-coref
[2]https://github.com/paragdakle/emailcoref
[3]https://catalog.ldc.upenn.edu/LDC2013T19

[4]http://rali.iro.umontreal.ca/rali/?q=en/wikicoref
[5]Hyperparameter values are provided in Appendix A.

Model	B^3	MUC	CEAFE	Avg. F1	#IC
LitBank (Bamman et al., 2020)					
Baseline	72.7	88.5	76.7	79.30	26
+ ET (orig)	**74.1**	**89.2**	**77.5**	**80.26**	**5**
+ ET (com)	73.3	89.1	77.5	79.97	13
EmailCoref (Dakle et al., 2020)					
Baseline	72.8	84.5	62.7	73.33	9
+ ET (orig)	**74.9**	86.7	**66.9**	**76.17**	**0**
+ ET (com)	74.8	**86.8**	66.7	76.10	2
WikiCoref (Ghaddar and Langlais, 2016)					
Baseline	70.7	80.0	54.7	68.45	133
+ ET (orig)	73.0	**82.8**	**58.3**	**71.35**	**70**
+ ET (com)	**73.4**	80.7	55.1	69.71	94
OntoNotes (Pradhan et al., 2012)					
Baseline	82.3	90.8	77.1	83.36	60
+ ET (orig)	84.4	**92.9**	**79.9**	**85.76**	**44**
+ ET (com)	**84.6**	92.6	79.8	85.68	46

Table 2: CR results with entity-type information (gold mentions). **IC = I**mpure **C**lusters.

5.1 Performance with Original Types

In order to establish an upper bound for improvement through introduction of type information, our first experiment leverages the original list of entity-types annotated in different corpora (**+ ET (orig)**), using the gold standard labels for types. Inclusion of entity-type information improves over the baseline for all CR datasets. Table 2 presents the performance of the baseline model and the model with entity-type information.

We find that entity-type information gives a boost of 0.96 Avg. F1 ($p < 0.01$) on LitBank which is the new state-of-the-art score with gold-mentions. This suggests that type information is helpful for CR on LitBank despite the heavily skewed distribution of entity-types in this corpus. Similarly, type information also benefits Email-Coref and WikiCoref resulting in an absolute improvement of 1.67 and 2.9 Avg. F1 points respectively ($p < 0.01$). We also see a 2.4 Avg. F1 improvement ($p < 0.01$) on OntoNotes, the largest dataset in this study. This suggests that explicit access to type information is beneficial all over the board, despite the use of contextual representations which have been claimed to model real-world facts and relationships (Petroni et al., 2019).

Ablation Results: To understand the contribution of the inclusion of type information to improve mention representation (+ET-self) and type consistency check between candidate mentions (+ET-

Dataset	Baseline	+ ET-self	+ ET-cross	+ ET (orig)
LitBank	79.30	79.97	79.76	**80.26**
EmailCoref	73.33	73.98	74.32	**76.17**
WikiCoref	68.45	70.89	70.71	**71.35**
OntoNotes	83.36	85.55	85.69	**85.76**

Table 3: Ablation results (**ET = ET-self + ET-cross**).

cross), we perform an ablation study (Table 3). We find that both components consistently provide significant performance boosts over the baseline. However, their combination (+ET) performs the best across all datasets.

5.2 Performance with Common Types

The previous experiment leverages the original entity-types assigned by dataset annotators. Due to the differences in domain and annotation guidelines among these datasets, the annotators introduce several domain-specific entity types (e.g. DIGital, Work Of Art etc.) apart from the common four (PERson, ORGanization, LOCation, FACility) that are often used in the Named Entity Recognition literature (Tjong Kim Sang, 2002). The former can prove to be much more difficult to obtain/ learn due to dearth of relevant data. Therefore, to assess the worth of using a common entity-type list for all datasets, we map the original types (Table 1) to the above-mentioned four common types. [6] Categories that do not map to any common type are assigned *Other*. **+ET (com)** rows in Table 2 show the results for this experiment. Models trained with common types as features perform worse than **+ET (orig)** which was expected as several original types are now clubbed into a single category (e.g. LAW -> OTHER, LANG -> OTHER) thus somewhat reducing the effectiveness of the feature. One surprising observation is the small difference between the performance on OntoNotes dataset, despite the fact that the number of type categories reduce from 18 + Other (**+ET (orig)**) to 4 + Other (**+ET (com)**). This could either be because (1) the entities with corpus-specific types occur less frequently in Ontonotes, or (2) the baseline model does a good job in resolving them. Further research is required to understand this case which is out of scope for this work.

---

[6]We provide the mapping between the original types and the common types in Appendix B.
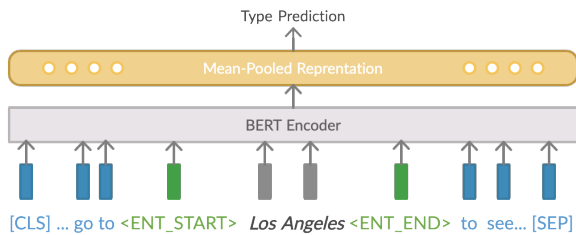
Figure 2: Type-prediction model.

5.3 # Impure Clusters (#IC)

Our hypothesis around the use of entity-types was to provide additional information to the model that could be leveraged to minimize errors due to type mismatch in CR. To evaluate if the F1 score improvements achieved by **+ET** models are because of fewer type mismatch errors, we report the number of coreference clusters detected by the model that contain at least one element with a type that is different from the others in the cluster. Since all of the datasets used in this work only consider identity coreferences (Recasens et al., 2011) - with potentially varied definitions of identity (Bamman et al., 2020; Pradhan et al., 2012) - where the mention is a linguistic "re-packaging" of its antecedent, this measure makes sense. As shown in Table 2, the models that score lower on the impurity measure get a higher Avg F1. This suggests that the aggregate performance improvements are at least partly due to the better mention-mention comparison in **+ET** systems.

6 Predicted Types

Results shown in the previous section assume the presence of gold standard types during training as well as inference, which is often impractical in the real-world. Most of the new samples that a CR model would encounter would not include type information about the candidate mentions. Therefore, we set up an additional experiment to gauge the benefits of type information using predicted types. We introduce a baseline approach to infer the type of the mentions and then use these predictions in the +ET models, in place of the gold types, for coreference resolution.

6.1 Type Prediction Model

Given the mention and its immediate context, i.e. the sentence it occurs in ($S = ..., c_{-2}, c_{-1}, \mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_n}, c_1, c_2, ...$), we add markers <ENT_START>/ <ENT_END>

before/ after the beginning/ ending of the mention in the sentence. The new sequence ($S' = ..., c_{-2}, c_{-1}$, <ENT_START>, $\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_n}$, <ENT_END>, $c_1, c_2, ...$) is tokenized using BERT tokenizer and passed through the BERT encoder. The output from which is then mean-pooled and passed through a fully-connected layer for classification. This architecture is motivated from (Soares et al., 2019) who show that adding markers around entities before passing the sentence through BERT performs better for relation extraction.

6.2 Experiments and Results

Type Prediction: Our final evaluation of the use of types in coreference is perhaps the most important one as it uses predicted types rather than annotated types, thus demonstrating that the benefits can be achieved in practice. Here we use the Type Prediction Model described just above. We limit the length of the input sequence to 128 tokens and use BERT-base-cased model for our type-prediction experiments. We perform a five-fold cross-validation to predict the type for each mention in the dataset. Since all four datasets suffer from class-imbalance, we report both Macro F1 score as well as the accuracy for the model. The model is trained for 20 epochs, with early-stopping (patience = 10), and is fine-tuned on the development set for Macro F1 to give more importance to minority type categories. We do not consider NA as a separate class during type prediction for WikiCoref and OntoNotes. For evaluation of our type-prediction model, we ignore the mentions that do not have an associated gold type (NA) from the final numbers in Table 4.

As shown, our model performs well on Lit-Bank, EmailCoref, and Ontonotes due to their favorable size in terms of training samples for the BERT-based type predictor. WikiCoref, however, proves more challenging as the model only manages 38.0 Macro F1 points with original (orig) types and 45.0 with common types (com), portraying its lack of ability to learn minority type categories with less data. Furthermore, our model finds it easier to predict the common (com) set of types for each dataset as combining multiple corpus-specific types into one partially alleviates the problem of class-imbalance. In line with our expectation, the largest improvement due to common types is seen for OntoNotes where the prob-

Dataset	Macro F1	Accuracy					
(Types)	All	All	PRP (dem.)	PRP (pers.)	NP (len = 1)	NP (len = 2)	NP (len > 2)
LitBank (orig)	84.0	97.0	70.5 (207)	99.75 (6127)	96.5 (6477)	95.8 (5439)	94.03 (4443)
LitBank (com)	87.0	97.0	72.5 (207)	99.8 (6127)	97.2 (6477)	96.7 (5439)	94.8 (4443)
EmailCoref (orig)	83.0	91.0	77.1 (35)	83.8 (482)	93.6 (1635)	91.3 (484)	94.1 (1058)
EmailCoref (com)	85.0	92.0	65.7 (35)	84.6 (482)	93.5 (1635)	90.3 (484)	95.5 (1058)
WikiCoref (orig)	38.0	61.0	53.3 (75)	59.4 (340)	58.9 (1531)	55.1 (1223)	54.5 (1218)
WikiCoref (com)	45.0	74.0	82.7 (75)	56.5 (340)	64.3 (1531)	69.2 (1223)	68.7 (1218)
OntoNotes (orig)	74.0	95.0	80.2 (1698)	94.6 (8682)	96.2 (21589)	95.5 (13297)	94.0 (11951)
OntoNotes (com)	90.0	96.0	84.5 (1698)	94.3 (8682)	96.9 (21589)	96.4 (13297)	95.0 (11951)

Table 4: Performance of our BERT-based model for type-prediction. The last five columns show the accuracy (# samples) on demonstrative and personal pronouns, and noun phrases of different lengths. **PRP (dem.)** = Demonstrative Pronouns (this, that, it, these, those), **PRP (pers.)** = Personal Pronouns (she, he, they, me, you, we). The scores for WikiCoref and OntoNotes do not include mentions without an associated gold type (*NA*).

Dataset	Baseline	+ ET-pred (orig)	+ ET-pred (com)
LitBank	79.30	**79.58**	79.40
EmailCoref	73.33	**74.20**	**74.18**
WikiCoref	68.45	68.37	68.62
OntoNotes	83.36	**84.02**	83.65

Table 5: CR results with predicted types. Numbers in **bold** are significantly better ($p < 0.01$) than the baseline.

lem reduces from an 18-way classification to a 5-way classification.

Coreference Resolution: Each mention in the corpus occurs in the test-sets of the five-fold cross-validation type-prediction experiments exactly once. This allows us to infer the type of each mention using the model that is trained on a different subset of the dataset. These inferred types are used in the training and testing of the CR systems in a manner similar to the annotated types.

Empirically, we found that the above configuration performs better than using the **+ET** models trained with annotated types and testing with predicted types, as the former exposes the CR models to the noisy types during training thus allowing them to learn weights that take this noise into account. We report the results for both original (**+ ET-pred (orig)**) and common (**+ ET-pred (com)**) type categories on each dataset.

Table 5 shows the results for performance of the baseline and the type-informed models on the four datasets, where the types are inferred from the model described in Section 6.1. We find that the improvements from type-information persist across LitBank, EmailCoref, and OntoNotes despite the use of predicted types, but, quite expectedly, remain smaller than the improvements from the gold annotated types. Scores on WikiCoref

show no significant improvement over the baseline, which could be explained by the poor performance of the type prediction model on this dataset which reduces the potency of the feature for CR.

7 Discussion

7.1 Genres of OntoNotes

Table 6 shows the most frequently occurring entity-types for each of the genres in OntoNotes. In line with our intuition, we find that enity-type information helps the baseline in *bc*, *bn*, *wb*, and *mz* genres which have less skew in their entity-type distribution. Genres like *bc*, *bn* and *wb*, although dominated by PER entities, contain a substantial minority of other entity-types like ORG and GPE. Along the same lines, *mz* contains a majority of GPE entities but also enough entities with type PER and ORG to make type information a potentially useful feature for CR. However, two exceptions to this are the improved performance of **+ET (orig)** on *tc* (highest skew) and no significant improvement on *nw* (lowest skew). These findings prompt further research in the future.

7.2 Type Prediction: PRP vs NP

Entity coreference in discourse often takes the surface form of pronouns (PRP) (like she, they, that, it etc.) or noun phrases (NP) (like LA, John's brother etc.) In Table 4, we compare the performance of our type prediction model on different types of pronouns, and noun phrases of varying length. We find that the model does well in predicting types for personal pronouns (PRP (pers.)) like *she*, *he* and noun phrases (NP). However, it consistently underperforms on demonstrative pro-

Genre	Baseline	+ET (orig)	Most Frequent Entity-Types (Ratio)
tc	81.15	**85.97**	PER (0.839), GPE (0.083)
bc	79.66	**83.25**	PER (0.456), GPE (0.277), NORP (0.098), ORG (0.089)
nw	84.91	85.26	ORG (0.411), GPE (0.232), PER (0.161), DATE (0.132)
pt	84.7	85.66	-
bn	84.22	**87.46**	PER (0.484), GPE (0.237), ORG (0.132), DATE (0.052)
wb	80.57	**83.52**	PER (0.744), GPE (0.130), ORG (0.092)
mz	89.21	**91.92**	GPE (0.513), PER (0.273), ORG (0.198)

Table 6: Type Distribution (excluding OTHER) for mentions within different Ontonotes genres after clustering based type-propagation. Numbers in **bold** are significantly better than the baseline ($p < 0.01$). - represents that none of the mentions in the genre are annotated with types.

	Example
Gold	Hi Chris **(0)** , I **(1)** will have {Tamara Utsch} **(2)** reply to you **(0)** on your (0) refund - I **(1)** believe she **(2)** updated you **(0)** last week , but we **(3)** ’ll see if there ’s new news this week . Can you **(0)** please send me **(1)** {your **(0)** current home address} **(4)** so we **(3)** can send you **(0)** {an organizer} **(5)** ? Thanks , {Judy Perdomo} **(1)** {PricewaterhouseCoopers Calgary} **(3)**
Baseline	Hi Chris **(0)** , I **(1)** will have {Tamara Utsch} **(2)** reply to you **(0)** on your (0) refund - I **(1)** believe she **(2)** updated you **(0)** last week , but we **(3)** ’ll see if there ’s new news this week . Can you **(0)** please send me **(1)** {your **(0)** current home address} **(4)** so we **(3)** can send you **(0)** {an organizer} *(4)* ? Thanks , {Judy Perdomo} **(1)** {PricewaterhouseCoopers Calgary} *(6)*
+ ET (orig)	Hi Chris **(0)** , I **(1)** will have {Tamara Utsch} **(2)** reply to you **(0)** on your (0) refund - I **(1)** believe she **(2)** updated you **(0)** last week , but we **(3)** ’ll see if there ’s new news this week . Can you **(0)** please send me **(1)** {your **(0)** current home address} **(4)** so we **(3)** can send you **(0)** {an organizer} **(5)** ? Thanks , {Judy Perdomo} **(1)** {PricewaterhouseCoopers Calgary} **(3)**
+ ET (com)	Hi Chris **(0)** , I **(1)** will have {Tamara Utsch} **(2)** reply to you **(0)** on your (0) refund - I **(1)** believe she **(2)** updated you **(0)** last week , but we **(3)** ’ll see if there ’s new news this week . Can you **(0)** please send me **(1)** {your **(0)** current home address} **(4)** so we **(3)** can send you **(0)** {an organizer} **(5)** ? Thanks , {Judy Perdomo} **(1)** {PricewaterhouseCoopers Calgary} **(3)**
+ ET-pred (orig)	Hi Chris **(0)** , I **(1)** will have {Tamara Utsch} **(2)** reply to you **(0)** on your (0) refund - I **(1)** believe she **(2)** updated you **(0)** last week , but we **(3)** ’ll see if there ’s new news this week . Can you **(0)** please send me **(1)** {your **(0)** current home address} **(4)** so we **(3)** can send you **(0)** {an organizer} **(5)** ? Thanks , {Judy Perdomo} **(1)** {PricewaterhouseCoopers Calgary} *(6)*

Table 7: An example email from EmailCoref corpus. Numbers in round brackets denote the cluster number of the mention. Red denotes incorrect predictions.

nouns (PRP (dem.)) like *this*, *that*, and *it* across all datasets.

This reduced performance could be due to the fact that demonstrative pronouns do not contain any signal about the type of the entity they refer to. Therefore, the type prediction model has to solely rely on the context to make that decision. However, this is not the case with PRPs (pers.) and NPs where the mention string is usually a strong indicator of the type. This problem is worsened by the imbalance due to the small presence of PRP (dem.) mentions in difference CR datasets. Since, the model does not encounter enough PRPs (dem.), it might not be able to learn to give high importance to context in these cases.

This could be partially alleviated by creating a separate type-prediction path for PRP (dem.) where the mention span is masked before it is passed through the model. A model that is trained with masked mentions would focus more on the context for type prediction and thus could lead to better performance on PRPs (dem.).

One could also experiment with training the type-prediction model on all of the mentions across the four datasets. The common list of types introduced in this work would allow for the creation of a larger training-set that includes mentions from multiple corpora (including external NER datasets) which could provide enough signal for the model to better learn the common types for PRPs (dem.).

Both these approaches could further boost the

results for CR with predicted entity-types, ultimately, reducing the gap between the scores in Table 2 and 5. However, they are left as future work as they are out of scope for this paper.

7.3 Case Study

Table 7 provides an excerpt of an email from EmailCoref corpus. As shown, the baseline model predicts the coreference clusters for *an organizer* (DIG) and *PricewaterhouseCoopers Calgary* (ORG) incorrectly. For the former, the model mistakes it as a reference to *your current home address* (LOC) which is corrected by the entity-type aware models. For the latter, the baseline considers *PricewaterhouseCoopers Calgary (PCC)* as part of a new coreference cluster, even though it refers to the organization of the email's sender which was previously referred to as *we* in the email. Models with access to gold type information (+ET (orig) and +ET (com)) are able to make that connection.

+ET-pred (orig), however, is unable to cluster *PCC* correctly which could be due to the fact that the type-prediction model incorrectly classifies the type of *we* as PER rather than ORG. This could lead to the CR model considering *PCC* (ORG) as a new entity in the discourse rather than a postcedent of *we*. This example demonstrates that sentence-level context might not be sufficient in some cases for mention type-disambiguation. We intend to experiment with models that capture long-term context and leverage external knowledge in the future.

8 Conclusion

In this work, we show the importance of using entity-type information in neural coreference resolution (CR) models with contextualized embeddings like BERT. Models which leverage type information, annotated in the corpus, substantially outperform the baseline on four CR datasets by reducing the number of type mismatches in detected coreference clusters. Since, these datasets vary in number and categories of the types they define, we also experiment with mapping the original corpus types to four common types (PER, ORG, LOC, FAC) based on previous NER research that can be learnt more easily through large NER datasets. Models which use these common types perform slightly worse than original types but still show significant improvements over the baseline systems.

The presence of gold standard types during CR inference is unlikely in practice. Therefore, we propose a model that infers the type of a mention given the mention span and its immediate context to use along side the proposed CR approach. In our evaluation, we find that using types predicted by our model for CR still performs significantly better than the baseline, thus offering stronger evidence that type information holds the potential for practical improvements for CR.

References

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in english literature. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 44–54.

Eric Bengtson 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*, pages 294–303.

Shuang Chen, Jinpeng Wang, Feng Jiang, and Chin-Yew Lin. 2020. Improving entity linking by modeling latent entity type information. *arXiv preprint arXiv:2001.01447*.

H. H. Clark. 1975. Bridging. In *Proceedings of TIN-LAP*.

Kevin Clark and Christopher D Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262.

Parag Pravin Dakle, Takshak Desai, and Dan Moldovan. 2020. A study on entity resolution for email conversations. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 65–73.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proc. of EMNLP-IJCNLP*.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Abbas Ghaddar and Philippe Langlais. 2016. Wikicoref: An english coreference-annotated corpus of wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142.

Aria Haghighi 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*, pages 385–393.

Jerry R Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

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, Luke Zettlemoyer, and Daniel S Weld. 2019. Bert for coreference resolution: Baselines and analysis. 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 5807–5812.

Shalom Lappin and Herbert J Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561.

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.

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.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Ruslan Mitkov. 1999. *Anaphora resolution: the state of the art*. Citeseer.

Vincent Ng. 2017. Machine learning for entity coreference resolution: A retrospective look at two decades of research. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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)*, pages 2227–2237.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Sadat Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, et al. 2018. Anaphora resolution with the arrau corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 192–199.

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*, CoNLL '12, page 1–40, USA. Association for Computational Linguistics.

Marta Recasens, Eduard Hovy, and M Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905.

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.

Josef Steinberger, Massimo Poesio, Mijail A Kabadjov, and Karel Ježek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management*, 6(43):1663–1680.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020a. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020b. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Bonnie Lynn Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive processes*, 6(2):107–135.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Appendix

A   Hyperparameters

Hyperparameter	Value
BERT	base-cased
BERT weights	freeze
BiLSTM hidden dim	200
Type embedding size	20
FC-layer 1 size	150
FC-layer 2 size	150
Dropout	0.2

Table A1: Hyperparameter values for our model. We refer the reader to `https://github.com/dbamman/lrec2020-coref` for the implementation of the baseline model.

B   Original types to Common types

Corpus-level	Common
PER	PER
LOC	LOC
FAC	FAC
GPE	LOC
VEH	Other
ORG	ORG

Table A2: Litbank

Original	Common
PER	PER
ORG	ORG
LOC	LOC
DIG	Other

Table A3: EmailCoref

Original	Common
Organization	ORG
Person	PER
Corporation	FAC
Event	Other
Place	LOC
Thing	Other
OTHER	Other
NA	Other

Table A4: WikiCoref

We use four common types (PERson, LOCation, FACility, ORGanization) and *Other* in **+ET**

Original	Common
ORG	ORG
WORK_OF_ART	Other
LOC	LOC
CARDINAL	Other
EVENT	Other
NORP	Other
GPE	LOC
DATE	Other
PERSON	PER
FAC	FAC
QUANTITY	Other
ORDINAL	Other
TIME	Other
PRODUCT	Other
PERCENT	Other
MONEY	Other
LAW	Other
LANGUAGE	Other
NA	Other

Table A5: OntoNotes

(com) experiments. These types are annotated in most of the named-entity recognition datasets and therefore are easier to model and learn via machine learning approaches. Tables A2, A3, A4, A5 show the mapping between the original types of each coreference dataset used in our study to the reduced common types. The most drastic difference occurs for OntoNotes (19 -> 5) and WikiCoref (8 -> 5). *OTHER* type in WikiCoref is for freebase links that did not have an associated type stored in freebase, whereas *NA* is used for mentions which do not have a freebase link. For OntoNotes, *NA* refers to the mentions that did not get any type assigned to them even after the use of our cluster based type-propagation approach (explained in Section 4).