Fine-grained Information Status Classification Using Discourse Context-Aware BERT

Yufang Hou
IBM Research Europe, Ireland
yhou@ie.ibm.com

Abstract

Previous work on bridging anaphora recognition (Hou et al., 2013a) casts the problem as a subtask of learning fine-grained information status (IS). However, these systems heavily depend on many hand-crafted linguistic features. In this paper, we propose a simple discourse context-aware BERT model for fine-grained IS classification. On the ISNotes corpus (Markert et al., 2012), our model achieves new state-of-the-art performance on fine-grained IS classification, obtaining a 4.8 absolute overall accuracy improvement compared to Hou et al. (2013a). More importantly, we also show an improvement of 10.5 F1 points for bridging anaphora recognition without using any complex hand-crafted semantic features designed for capturing the bridging phenomenon. We further analyze the trained model and find that the most attended signals for each IS category correspond well to linguistic notions of information status.

1 Introduction

Information Structure (Halliday, 1967; Prince, 1981; Prince, 1992; Gundel et al., 1993; Lambrecht, 1994; Birner and Ward, 1998; Kruijff-Korbayová and Steedman, 2003) studies structural and semantic properties of a sentence according to its relation to the discourse context. Information structure affects how discourse entities are referred to in a text, which is known as *Information Status* (Halliday, 1967; Prince, 1981; Nissim et al., 2004). Specifically, information status (IS henceforth) reflects the accessibility of a discourse entity based on the evolving discourse context and the speaker's assumption about the hearer's knowledge and beliefs. For instance, according to Markert et al. (2012), *old* mentions[1] refer to entities that have been referred to previously; *mediated* mentions have not been mentioned before but are accessible to the hearer by reference to another *old* mention or to prior world knowledge; and *new* mentions refer to entities that are introduced to the discourse for the first time and are not known to the hearer before.

In this paper, we mainly follow the IS scheme proposed by Markert et al. (2012) and focus on learning fine-grained IS on written texts. A mention's semantic and syntactic properties can signal its information status. For instance, indefinite NPs tend to be *new* and pronouns are likely to be *old*. Moreover, referential patterns of how a mention is referred to in a sentence also affect this mention's IS. In Example 1, "Friends" is a bridging anaphor even if we do not know the antecedent (i.e., *she*); while the information status for "Friends" in Example 2 is *mediated/worldKnowledge*. Section 3.1 analyzes the characteristics of each IS category and the relations between IS and discourse context.

(1) [*She*]$_{antecedent}$ made money, but spent more. **Friends** pitched in.

(2) <u>Friends</u> are part of the glue that holds life and faith together.

In this work, we propose a simple yet effective discourse context-aware self-attention model based on BERT (Devlin et al., 2019) for fine-grained IS classification. We find that the sentence containing the target mention as well as the lexical overlap information between the target mention and the preceding

[1] A mention is a noun phrase which refers to a discourse entity and carries information status.

Proceedings of the 28th International Conference on Computational Linguistics, pages 6101–6112
Barcelona, Spain (Online), December 8-13, 2020

mentions are the most important discourse context when assigning IS for a mention. With the self-attention mechanism, our model can capture important signals within a mention and the interactions between the mention and its context. On the ISNotes corpus (Markert et al., 2012), our model achieves new state-of-the-art performance on fine-grained IS classification, obtaining a 4.8 absolute overall accuracy improvement compared to Hou et al. (2013a). More importantly, we also show an improvement of 10.5 F1 points for bridging anaphora recognition without using any sophisticated hand-crafted semantic features.

Furthermore, to gain additional insights into our model's predictions, we analyze the attention mechanisms of our trained model. We find that the most attended tokens for each IS category correspond well with linguistic features of information status. For instance, for the *old* IS category, the most attended token list includes pronouns such as "she", "her", and "it". While for the *new* category, the model pays more attention to indefinite determiners such as "a" and "an". Section 6 provides a detailed analysis of the attention map for each IS category.

To summarize, the main contributions of our work are as follows:

- We propose a simple and effective model for fine-grained IS classification. The model uses a novel approach for encoding information from the previous sentences along with the current sentence for IS classification.

- Our proposed model achieves new state-of-the-art results for IS classification and bridging anaphora recognition on the ISNotes corpus. Our model also achieves competitive results for fine-grained IS classification on the Switchboard dialogue IS corpus (Nissim et al., 2004) that uses a different IS scheme than the one in ISNotes. The processed datasets and code are publicly available at: `https://github.com/IBM/bridging-resolution`.

- We carry out ablation studies to understand the effectiveness of each component in our model. We further investigate the self-attention patterns in our model and find that the model does learn specific linguistic features for predicting information status.

2 Related Work

IS classification and bridging anaphora recognition. Bridging resolution (Hou et al., 2014; Hou et al., 2018) contains two sub tasks: identifying bridging anaphors (Markert et al., 2012; Hou et al., 2013a; Hou, 2016a) and finding the correct antecedents among candidates (Hou et al., 2013b; Hou, 2018a; Hou, 2018b; Hou, 2020). Most previous studies handle bridging anaphora recognition as part of IS classification problem. Markert et al. (2012) applied joint inference for IS classification on the ISNotes corpus but reported very low results on bridging recognition. Building on this work, Hou et al. (2013a) designed many linguistic features to capture bridging and integrated them into a cascading collective classification algorithm. This approach later was integrated into a pipeline for bridging resolution (Hou, 2016b; Hou et al., 2018). Differently, Hou (2016a) used an attention-based LSTM model based on GloVe vectors and a small set of features for IS classification. The author reported similar results as Hou et al. (2013a) regarding the overall IS classification accuracy but the result on bridging anaphora recognition is much worse than Hou et al. (2013a).

Rahman and Ng (2012) incorporated carefully designed rules into an SVM algorithm for IS classification on the Switchboard dialogue IS corpus (Nissim et al., 2004).[2] The authors first designed a rule-based system to assign IS classes to mentions on the basis of Nissim's IS annotation guidelines (Nissim et al., 2004). They then applied an SVM^{multiclass} algorithm for this task by combining the prediction from the rule-based system, the ordering of the rules as well as two lexical features.

Another work on IS classification was carried out by Cahill and Riester (2012). They assumed that the distribution of IS classes within sentences tends to have certain linear patterns, e.g., *old* > *mediated* > *new*. Under this assumption, they trained a CRF model with syntactic and surface features for fine-grained IS classification on the German DIRNDL radio news corpus (Riester et al., 2010). Recently, Rösiger (2019)

[2] It is worth noting that bridging antecedent information was not annotated in Switchboard. Also, bridging anaphora annotation in Switchboard includes non-anaphoric cases.

adapted eight rules from Hou et al. (2014) to recognize bridging anaphors and find their antecedents in the improved annotations of the extended DIRNDL corpus (Björkelund et al., 2014).

Different from the above-mentioned work, we do not use any complicated hand-crafted features, and our model improves the previous state-of-the-art results on both overall IS classification accuracy and bridging recognition by a large margin on the ISNotes corpus. Our model also achieves competitive results for fine-grained IS classification on Switchboard compared to the approach in Rahman and Ng (2012) that uses the Stanford coreference resolver and an SVM classifier that explores 18 carefully designed hand-crafted rules.

Fine-tuning with contextual word embeddings. Recent studies (Peters et al., 2018; Devlin et al., 2019) have shown that a range of downstream NLP tasks benefit from fine-tuning task-specific parameters with pre-trained contextual word representations. Our work belongs to this category and we fine-tune our model based on BERT representations (Devlin et al., 2019). The novelty of our approach is that we create a "pseudo sentence" for each mention that encodes the most effective local and global discourse context for predicting the mention's IS. The self-attention mechanism in Transformer's self-attention encoder (Ashish et al., 2017) allows our model to attend to both the context and the mention itself for clues that are helpful to predict the mention's IS.

Model probing. Recently, there have been a number of studies exploring the types of knowledge encoded in the BERT model. Jawahar et al. (2019) found that internal vector representations in BERT encode rich linguistic information, with surface information at the bottom layers, syntactic information in the middle layers, and semantic information at the top layers. Clark et al. (2019) showed that certain attention heads in BERT correspond well to the linguistic knowledge of syntax and coreference. In our work, we demonstrate that the attention patterns in our trained model embed linguistic notions of information status.

3 Approach

3.1 Information Status and Discourse Context

The IS scheme proposed by Markert et al. (2012) adopts three major course-grained IS categories (*old*, *new*, and *mediated*) from Nissim et al. (2004) and distinguishes six subcategories for *mediated*. Below we provide a brief description for the eight fine-grained IS classes in ISNotes.

Old mentions are coreferent with the already introduced entities. *New* mentions are entities that have not been introduced into the discourse and the hearer/reader cannot infer them from either previously mentioned entities or general world knowledge. *Mediated* mentions are discourse-new and hearer-old (Prince, 1992). They have not been introduced into the discourse before but are accessible to the hearer by reference to another mention or to prior world knowledge.

Among the *mediated* category, *Mediated/worldKnowledge* mentions are generally known to the hearer. This category contains mostly proper names. *Mediated/syntactic* mentions are syntactically linked to other *old* or *mediated* mentions, such as "[[their]$_{old}$ father]$_{m/syntactic}$" or "[a war in [Africa]$_{mediated}$]$_{m/syntactic}$". *Mediated/aggregate* mentions are coordinated NPs where at least one element is *old* or *mediated*, such as "[[U.S.]$_{mediated}$ and [Canada]$_{mediated}$]$_{m/aggregate}$". *Mediated/function* mentions refer to a value of a previously explicitly mentioned function and this function needs to be able to rise or fall (e.g., **6 cents** in Example 3). *Mediated/comparative* mentions usually contain a premodifier to indicate that this entity is compared to another preceding entity (antecedent) (e.g., **further attacks** in Example 4). Finally, *Mediated/bridging* mentions are associative anaphors that link to previously introduced related entities/events (e.g., **Friends** in Example 1).

(3) In trading on the American Stock Exchange, Delmed's price [went down]$_{function}$ **6 cents**.

(4) [*The cyber attacks*]$_{antecedent}$ were followed by **further attacks** on ZDNet.com, a news portal.

We characterize the linguistic factors that affect a mention's IS into three categories: mention properties, local context, as well as previous context. Table 1 lists the definitions for these IS categories and

			Factors affecting IS		
	Description	Example	Mention Properties	Local Context	Previous Context
old	coreferent with an already introduced entity	*he*, *the president*	✓	✓	✓
m/worldKnow.	generally known to the hearer	*Francis*, *the pope*	✓	✓	
m/syntactic	syntactically linked to other *old* or *mediated* mentions	*their father* *a war in Africa*	✓		
m/aggregate	coordinated NPs where at least one element is *old* or *mediated*	*U.S. and Canada* *he and his son*	✓		
m/function	refer to a value of a previously explicitly mentioned rise/fall function	(the price went down) *6 cents*	✓	✓	
m/comparative	usually contain a premodifier to indicate that this entity is compared to another entity	*another law* *further attacks*	✓	✓	✓
m/bridging	associative anaphors which link to previously introduced related entities/events	*the price* *the reason*	✓	✓	✓
new	introduced into the discourse for the first time and not known to the hearer before	*a reader* *politics*	✓	✓	✓

Table 1: Information status categories and their main affecting factors. “Local context” means the sentence s which contains the target mention, “Previous context” indicates all sentences from the discourse which occur before s.

summarizes the main affecting factors for each IS class. Note that we analyze the main affecting factors for each IS class based on their definitions.

As described in Section 1, a mention’s internal syntactic and semantic properties can signal its IS. For instance, a mention containing a possessive pronoun modifier is likely to be *mediated/syntactic* (e.g., *their father*); and a *mediated/comparative* mention often contains a premodifier indicating that this entity is compared to another preceding entity (e.g., *further attacks*).

In addition, for some IS classes, the “local context” (the sentence s which contains the target mention) and “previous context” (sentences from the discourse which precede s) play an important role when assigning IS to a mention. Example 1 and Example 2 in Section 1 demonstrate the role of the local context for IS. In Example 1, the referential patterns in the local context indicate that “Friends” is a bridging anaphor,[3] whereas “Friends” in Example 2 is a generic NP.

Sometimes we need to look at the previous context when deciding IS for a mention. In Example 5, without looking at the previous context, we tend to think the IS for “Poland” in the second sentence is *mediated/WorldKnowledge*. Here the correct IS for “Poland” is *old* because it is mentioned before in the previous context.

(5) [Previous context:] In **Poland**, only 4% of all investment goes toward making things farmers want; in the West, it is closer to 20%.
[Local context:] A private farmer in **Poland** is free to buy and sell land.

3.2 IS Classification with Discourse Context-Aware Self-Attention

To account for the different factors described in the previous section when predicting IS for a mention, we create a novel “pseudo sentence” for each mention and apply the multi-head self-attention encoder (Ashish et al., 2017; Devlin et al., 2019) to this sentence.

Figure 1 depicts the high-level structure of our model. The pseudo sentence consists of five parts: previous overlap_info, local context, the delimiter token “[SEP]”, the content of the target mention, and the IS prediction token “[CLS]”. The previous overlap_info part contains two tokens, which indicate whether the target mention has the same string/head with a mention from the preceding sentences. And the local context is the sentence containing the target mention.

The final prediction is made based on the hidden state of the prediction token “[CLS]”. In principle, this is similar to BERT’s “[CLS]a[SEP]b” framework, in which the special classification token (*[CLS]*)

[3] Clark (1975) uses necessary, probable and inducible parts/roles to distinguish different types of bridging and argues that only in the first case the antecedent triggers the bridging anaphor in the sense that we already spontaneously think of the anaphor when we read/hear the antecedent. In the probable/inducible cases, the bridging anaphora accommodates itself into the context and is induced by the need for an antecedent. Section 4.3 illustrates this using a wug-test example.

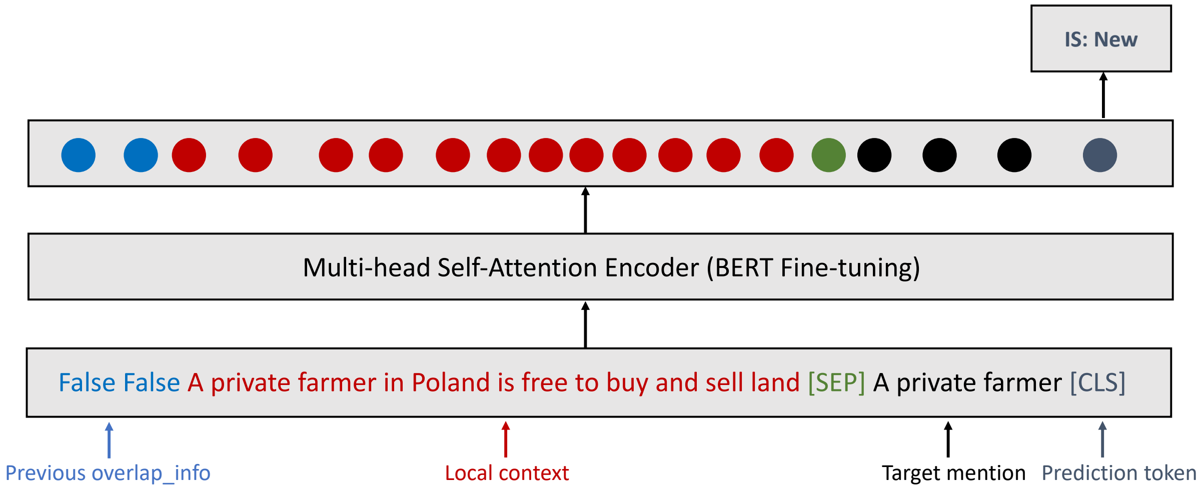
IS: New
Multi-head Self-Attention Encoder (BERT Fine-tuning)
False False A private farmer in Poland is free to buy and sell land [SEP] A private farmer [CLS]
Previous overlap_info
Local context
Target mention
Prediction token

Figure 1: Fine-grained IS classification with discourse context-aware self-attention.

is added to every sequence as the first token and its hidden state is used as the aggregate sequence representation for classification tasks. The novelty of our work is that we design the structure of *a* and *b* in a way that embeds the most indicative information to predict a mention's information status. During the training stage, the mechanism of multi-head self-attention helps the model to learn the important cues from both the mention itself and its discourse context when predicting IS.

There are other ways to encode a mention's context information. For instance, one could try to add more previous sentences in the local context or replace the current *previous overlap_info* with all previous sentences. In practice, we found that the current configuration yields the best results on the ISNotes and Switchboard IS corpora. In particular, we notice that using all previous sentences as the discourse context significantly decreases the results for IS classification.[4] This is in line with the observation from Joshi et al. (2019) that modeling the longer context in BERT provides no improvement for coreference resolution.

3.3 Model Parameters

We use the vanilla BERT (Devlin et al., 2019) for our experiments. We initialize our model using pre-trained BERT contextual embeddings, which is trained on top of the BookCorpus (800M words) and English Wikipedia (2,500M words). We then fine-tune the model for 3 epochs with the learning rate of $3e - 5$ and a batch size of 32. During training and testing, the max token size of the pseudo sentence is set as 128.[5]

4 Experiments on ISNotes

4.1 Experimental Setup

We perform experiments on the ISNotes corpus (Markert et al., 2012), which contains 10,980 mentions annotated for information status in 50 news texts taken from the Wall Street Journal portion of the OntoNotes corpus (Weischedel et al., 2011). Table 2 shows the IS distribution in ISNotes.

Following Hou et al. (2013a), all experiments are performed via 10-fold cross-validation on documents. On each testing fold, the model is trained on the other nine folds. The hyper-parameters of 3 epochs and the learning rate $3e - 5$ were fixed during all training processes. We report overall accuracy as well as precision, recall and F-score per IS class. In the following, we describe the baselines as well as our model with different settings.

[4] When the pseudo sentence is longer than 512 tokens, we use a sliding window of 100 tokens to account for all previous context of a mention.

[5] The length of 128 tokens covers all the cases in ISNotes. In practice, we truncate the local sentence if the whole pseudo sentence is longer than 128 tokens.

Mentions	10,980	
old	3,237	29.5%
mediated	3,708	33.8%
syntactic	1,592	14.5%
world knowledge	924	8.4%
bridging	663	6.0%
comparative	253	2.3%
aggregate	211	1.9%
func	65	0.6%
new	4,035	36.7%

Table 2: IS distribution in ISNotes.

	baselines									*this work*					
	collective			*cascade collective*			*incremental LSTM*			*self-attention with*			*self-attention with*		
	Hou et al.(2013)			Hou et al.(2013)			Hou (2016)			*BERT$_{BASE}$*			*BERT$_{LARGE}$*		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
old	84.4	86.0	85.2	82.2	87.2	84.7	85.4	84.9	85.2	87.8	90.5	89.1	88.4	90.0	**89.2**
m/worldKnow.	67.4	77.3	72.0	67.2	77.2	71.9	67.1	74.5	70.6	74.9	77.8	76.3	77.7	79.5	**78.6**
m/syntactic	82.2	81.9	82.0	81.6	82.5	82.0	80.8	81.9	81.4	82.9	79.5	81.1	83.7	81.1	**82.4**
m/aggregate	64.5	79.5	71.2	63.5	77.9	70.0	67.8	84.6	75.3	76.8	74.7	75.7	80.1	79.3	**79.7**
m/function	67.7	72.1	69.8	67.7	72.1	69.8	64.6	76.4	70.0	35.9	62.2	45.5	73.4	85.5	**79.0**
m/comparative	81.8	82.1	82.0	86.6	78.2	82.2	77.9	83.1	80.4	88.1	84.8	86.4	90.5	86.7	**88.6**
m/bridging	19.3	39.0	25.8	44.9	39.8	42.2	15.7	32.3	21.1	43.3	49.6	46.2	51.0	54.5	**52.7**
new	86.5	76.1	81.0	83.0	78.1	80.5	87.2	74.8	80.5	85.7	82.5	84.1	86.6	85.2	**85.9**
acc	78.9			78.6			78.6			82.0			**83.7**		

Table 3: Results of the discourse context-aware self-attention model compared to the baselines on ISNotes. Bolded scores indicate the best performance for each IS class. The improvements of *self-attention with BERT$_{BASE}$* and *self-attention with BERT$_{LARGE}$* over the baselines are statistically significant at $p<0.01$ using randomization test.

***collective* (baseline 1).** Hou et al. (2013a) applied collective classification to account for the linguistic relations among IS categories. They explored a wide range of features (34 in total), including a large number of lexico-semantic features (for recognizing bridging) as well as a couple of surface features and syntactic features.

***cascaded collective* (baseline 2).** This is the cascading minority preference system for bridging anaphora recognition from Hou et al. (2013a).

***incremental LSTM* (baseline 3).** This is the attention-based LSTM model proposed by Hou (2016a). The model uses one-hot vectors to encode IS classes and predicts information status for all mentions of a document from left to right incrementally.

***self-attention with BERT$_{BASE}$*.** We fine-tune BERT$_{BASE}$ on the pseudo sentences described in Section 3. The model has 12 transformer blocks, 768 hidden units, and 12 self-attention heads.

***self-attention with BERT$_{LARGE}$*.** We fine-tune BERT$_{LARGE}$ on the pseudo sentences described in Section 3. The model has 24 transformer blocks, 1024 hidden units, and 16 self-attention heads.

4.2 Results and Discussion

Table 3 shows the results of our models compared to the baselines. Our best model *self-attention with BERT$_{LARGE}$* improves over all baselines by a large margin on all IS categories. It achieves an overall accuracy of 83.7% on fine-grained IS classification, obtaining a 4.8 and 5.1 absolute improvements in accuracy over the two strong baselines (*collective* and *cascade collective*), respectively.

	self-attention with BERT$_{LARGE}$			*self-attention wo target mention*			*self-attention wo local context*			*self-attention wo pre_overlap_info*			*self-attention wo local context pre_overlap_info*		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
old	88.4	90.0	**89.2**	78.2	75.8	77.0	87.1	90.8	88.9	80.8	87.4	84.0	74.8	84.9	79.5
m/worldKnow.	77.7	79.5	**78.6**	15.9	37.7	22.4	75.9	77.6	76.7	70.1	64.1	67.0	64.7	56.5	60.3
m/syntactic	83.7	81.1	82.4	23.8	42.8	30.6	85.1	81.2	83.1	85.6	81.4	**83.4**	84.6	80.3	82.4
m/aggregate	80.1	79.3	79.7	15.2	38.1	21.7	84.4	78.8	**81.5**	77.3	81.5	79.3	78.7	78.7	78.7
m/function	73.4	85.5	79.0	43.8	41.8	42.7	62.5	62.5	62.5	75.0	90.6	**82.1**	65.6	65.6	65.6
m/comparative	90.5	86.7	**88.6**	6.7	31.5	11.1	90.5	86.7	**88.6**	89.7	86.0	87.8	88.9	84.3	86.5
m/bridging	51.0	54.5	**52.7**	4.7	38.3	8.3	43.7	48.9	46.2	51.9	52.5	52.2	40.3	46.4	43.1
new	86.6	85.2	**85.9**	80.9	53.7	64.5	85.1	82.6	83.8	86.5	84.6	85.5	85.1	80.3	82.6
acc		**83.7**			58.5			82.3			81.1			77.4	

Table 4: Ablation experiments results of *self-attention with BERT$_{LARGE}$* for fine-grained IS classification. Bolded scores indicate the best performance for each IS class. The differences between *self-attention with BERT$_{LARGE}$* and other variations are statistically significant at p<0.01 using randomization test.

It is worth noting that recognizing bridging anaphora is a challenging task (Markert et al., 2012). Hou et al. (2013a) proposed a lot of discourse structure, lexico-semantic and genericity detection features to capture the phenomenon. Their best model for bridging anaphora recognition (*cascade collective*) achieves an F-score of 42.2. Overall, our model *self-attention with BERT$_{LARGE}$* achieves the new state-of-the-art performance for this task with an F-score of 52.7 without resorting to any hand-crafted sophisticated semantic features. By comparing the confusion matrices of *cascade collective* and *self-attention with BERT$_{LARGE}$*, we find that the highest proportion of recall errors of bridging recognition in *cascade collective* is due to the fact that a lot of bridging anaphors are misclassified as *new*. This can be explained as the syntactic form of many new mentions and bridging anaphors are the same (see Example 1 and Example 2), the lexico-semantic features in *cascade collective* only pick up on certain types of bridging. In addition, most precision errors in *cascade collective* are *new* and *old* mentions being misclassified as *m/bridging*. Both these recall and precision errors are less frequent in *self-attention with BERT$_{LARGE}$*. It seems that our model does capture properties of bridging anaphora better by only looking at a mention and its interactions with the surrounding context.

4.3 Ablations

To better understand the impact of different components in our model, we carry out an ablation experiment. We remove *target mention*, *local context*, *previous overlap_info*, as well as all context information (*local context* + *previous overlap_info*) from our best model *self-attention with BERT$_{LARGE}$*, respectively. Table 4 reports the results of different configurations for our model.

Surprisingly, the model considering only the content of mentions (see the last column of Table 4) achieves competitive results as the baseline *cascade collective* which explores many hand-crafted linguistic features. Also it outperforms the three baselines on several IS categories (*m/syntactic*, *m/aggregate*, *m/comparative*, *m/bridging* and *new*). In Section 3.1, we analyze that *m/syntactic* and *m/aggregate* are often signaled by mentions' internal syntactic structures, and that the semantics of certain premodifiers is a strong signal for *m/comparative*. The improvements on these categories show that our model can capture the semantic/syntactic properties of a mention when predicting its IS.

Among all three components, it seems that the content of mentions has the most impact on the overall results, while the local context has the least impact. Furthermore, we find that *local context* and *previous overlap_info* have different impacts on IS classes. More specifically, we notice that *m/bridging*, *m/function* and *new* benefit most from *local context*, whereas *old* and *m/worldKnowledge* benefit most from *previous overlap_info*. This may seem counter-intuitive for *m/bridging* and *m/worldKnowledge*, as one expects that *m/bridging* should benefit more from the previous context and *m/worldKnowledge* is a local phenomenon. For *m/worldKnowledge*, this is explained by the fact that the system without previous context information (*self-attention wo pre_overlap_info*) wrongly predicts a lot of *old* mentions as *m/worldKnowledge*, as illustrated in Example 5.

	SVM + Stanford Coreference Rahman and Ng (2012)			*self-attention with* $BERT_{LARGE}$		
	R	P	F	R	P	F
old/ident	75.8	64.2	69.5	81.9	82.5	82.2
old/event	2.4	31.8	4.5	73.2	71.5	72.3
old/general	87.8	92.7	90.2	96.1	97.1	96.6
old/generic	39.9	85.9	54.5	77.4	73.7	75.5
old/ident_generic	47.2	44.8	46.0	74.8	72.6	73.7
old/relative	99.0	37.5	54.4	97.4	95.4	96.4
med/general	84.0	72.2	77.7	85.3	74.6	79.6
med/bound	2.7	40.0	5.1	29.7	43.4	35.3
med/part	73.2	96.8	83.3	51.8	54.7	53.2
med/situation	68.0	97.7	80.2	22.4	38.7	28.4
med/event	46.3	100.0	63.3	16.2	22.9	19.0
med/set	88.4	86.0	87.2	78.1	71.9	74.9
med/poss	90.5	97.6	93.9	82.5	77.8	80.1
med/func_value	88.1	85.9	87.0	40.0	10.5	16.7
med/aggregation	83.8	93.9	88.6	56.7	46.0	50.7
new	90.4	83.6	86.9	67.9	76.9	72.1
acc	78.7			**79.3**		

Table 5: Results of our discourse context-aware self-attention model compared to the *SVM + Stanford Coreference* model (Rahman and Ng, 2012) on the Switchboard dialogue IS corpus for fine-grained IS classification.

For *m/bridging*, the big impact of the local context corresponds to Hou et al. (2013a)'s observation that some bridging can be indicated by referential patterns without world knowledge about the anaphor/antecedent NPs. For instance, in the following sentence, "*The blicket couldn't be connected to the dax. **The wug** failed.*", the mention "***The wug***" is likely a bridging anaphor, although we do not know the antecedent.[6] Similarly, Clark (1975) distinguishes between bridging via necessary, probable, and inducible parts/roles. He states that only in the first case does the antecedent trigger the bridging anaphor in the sense that we already spontaneously think of the anaphor when we read/hear the antecedent. In the probable/inducible cases, bridging anaphora accommodates itself into the context and is induced by the need for an antecedent.

In addition, we also tested whether a broader local context can help us to detect bridging better. In the ISNotes corpus, 26% of bridging anaphors have the antecedents from the same sentence, and 77% of anaphors have antecedents occurring in the same or up to two sentences prior to the anaphor. In practice, we tried to add the previous k sentences ($k = 1$ and $k = 2$) into the current local context but found that the overall results for bridging in both settings are similar to the current one.

5 Experimental Results on the Switchboard IS Corpus

In this section, we apply our discourse context-aware self-attention model to the Switchboard dialogue IS corpus (Nissim et al., 2004). The corpus contains around 63k mentions annotated with IS types (i.e., *old*, *mediated*, and *new*) and subtypes. Note that the IS scheme in this corpus is different from the one in ISNotes in terms of fine-grained IS classes. In general, bridging in this corpus includes non-anaphoric, syntactically linked part-of and set-member relations (e.g., *the house's door*), as well as comparative anaphors that are marked by surface indicators such as "other" or "different".[7] Nevertheless, we think the

[6] We thank an anonymous reviewer for bringing up this example in one of our previous work (Hou et al., 2013a). We also thank another anonymous reviewer of this work for pointing out that "***The wug***" could also be an epithet.

[7] We refer the readers to Nissim et al. (2004) for more details of fine-grained IS categories in the Switchboard dialogue IS corpus.

IS class	Most attended tokens
old	the, pre_overlap2 = NA, pre_overlap1 = NA, pre_overlap2 = yes, pre_overlap1 = yes, it, her, she, that, they
m/worldKnow.	pre_overlap1 = no, pre_overlap2 = no, the, month, year, and, of, to, this, said
m/syntactic	pre_overlap1 = no, the, pre_overlap2 = no, of, 's, her, in, its, pre_overlap2 = yes, to
m/aggregate	and, the, or, pre_overlap1 = no, pre_overlap2 = yes, her, oil, units, of, -
m/function	%, units, pre_overlap2 = yes, pre_overlap1 = no, to, 8, fell, 5, 243, million
m/comparative	pre_overlap1 = no, more, pre_overlap2 = no, other, pre_overlap2 = yes, higher, companies, some, of, that
m/bridging	pre_overlap1 = no, the, pre_overlap2 = no, a, in, friends, year, demand, production, to
new	pre_overlap1 = no, pre_overlap2 = no, the, a, an, of, to, -, magazines, but

Table 6: Top ten most attended tokens for each IS class in *self-attention with BERT$_{LARGE}$* trained on ISNotes.

three main linguistic factors affecting a mention's IS analyzed in Section 3.1 still hold in the dialogue domain.

Following Rahman and Ng (2012), we split the dataset into a training set containing 117 dialogues and a testing set containing 30 dialogues. We train our model *self-attention with BERT$_{LARGE}$* on the training dataset using the parameters described in Section 3.3. Table 5 lists the results of our model compared to Rahman and Ng (2012)'s system, which is an SVM^{multiclass} model based on predictions from a rule-based system and the Stanford Deterministic Coreference Resolution System (Lee et al., 2011). The rule-based system consists of 18 hand-crafted rules for assigning IS subtypes to mentions. Some rules are based on the lexicon relations encoded in WordNet and FrameNet.

Note that the results of the two systems in Table 5 are not directly comparable due to the different splits of the training/testing datasets.[8] Neverthless, our model *self-attention with BERT$_{LARGE}$* achieves competitive performance compared to Rahman and Ng (2012)'s system in terms of the overall accuracy. In general, it seems that our model is better at predicting *old* mentions. We also checked the confusion matrix and found that the low results for *med/situation*, *med/event* and *med/func_value* is due to the fact that our model cannot distinguish these three categories from *med/set*.

6 Attention to Linguistic Features

In order to gain additional insights into our model's predictions, we analyze the attention maps in our best model (*self-attention with BERT$_{LARGE}$*) that is trained on ISNotes. We aim to check to what extent the most attended tokens correspond to the linguistic features for each IS class.

Specifically, we randomly choose one fold and apply the trained model to the testing dataset. Since the "[CLS]" token is used for prediction, we analyze the attention weights assigned to other tokens from "[CLS]" for each testing instance. The weight of each token is normalized by the sequence length. The final attended score for each token is calculated by aggregating normalized attended weights across all testing instances in all 16 heads from the last layer. This is because previous work suggests that the last layer usually encodes the task-specific features in fine-tuning (Kovaleva et al., 2019).

Table 6 lists the top ten most attended tokens for each IS class. We exclude the separator tokens ([CLS]/[SEP]) and two punctuation tokens (comma and period) from the list, as suggested by Clark et al. (2019) that these tokens are heavily attended in deep heads and might be used as a no-op for attention heads. Note that *pre_overlap1* and *pre_overlap2* are the two tokens that indicate whether the target mention has the same string/head with a mention from the preceding sentences. Both can have a value of "yes", "no", or "NA". Following Markert et al. (2012), "NA" means "non-applicable" and is mainly used for pronouns.

[8] Rahman and Ng (2012) reported that their testing dataset contains 30 dialogues, but the split of the training/testing datasets is not publicly available.

We notice that a lot of the attended tokens in Table 6 correspond well with the linguistic features for each IS class. For *old* mentions, the model attends to pronouns and signals that indicate string overlap, while for *new* mentions, the model attends to tokens that indicate string non-overlap and the indefinite determiners “a/an”.

It is interesting to note that the model seems to learn the internal syntactic/semantic structure for a few IS classes. For instance, “of” and “’s” are strong signals for *m/syntactic* mentions that have a prepositional structure or a possessive structure. Also *m/aggregate* mentions usually contain the tokens “and/or” that indicate the coordination structure. Similarly, for *m/comparative* category, the model learns to focus on a few premodifiers (e.g., “more”, “other”, and “higher”) that indicate the comparison between two entities.

Finally for *m/function* mentions, the model learns to mostly focus on numbers. Surprisingly, the model also learns to attend the verb “fell”, which corresponds well with the definition of this IS class (see Section 3.1). For the most difficult category *m/bridging*, it seems that the model attends to some relational nouns (e.g., “friends” or “demand”) that are likely used as bridging anaphors.

7 Conclusions

We propose a simple discourse context-aware self-attention model for IS classification based on the BERT fine-tuning framework. We cast the IS classification problem as a sentence classification task by creating a novel “pseudo sentence” for each mention. We design the “pseudo sentence” based on the linguistic intuitions about IS and it contains most indicative context information to predict a mention’s information status. Such design allows the model to capture both clues from the mention and its context when predicting IS.

Our model does not contain any complex hand-crafted semantic features and achieves the new state-of-the-art results for IS classification and bridging anaphora recognition on ISNotes that contains written news articles. In another domain that consists of conversational dialogues (Switchboard), our model also achieves competitive performance for fine-grained IS classification compared to previous work (Rahman and Ng, 2012).

Finally, in order to better understand our model’s predictions, we probe our best model (*self-attention with* $BERT_{LARGE}$) on ISNotes. We find that our model learns to pay more attention to signals that correspond well to the linguistic features of each IS class.

Acknowledgments

The author would like to thank Charles Jochim for proofreading the paper and the anonymous reviewers for their valuable feedback.

References

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Lukasz, and Polosukhin Illia. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017),* pages 1–11. MIT Press.

Betty J. Birner and Gregory Ward. 1998. *Information Status and Noncanonical Word Order in English.* John Benjamins, Amsterdam, The Netherlands.

Anders Björkelund, Kerstin Eckart, Arndt Riester, Nadja Schauffler, and Katrin Schweitzer. 2014. The extended DIRNDL corpus as a resource for coreference and bridging resolution. In *Proceedings of the 9th International Conference on Language Resources and Evaluation,* Reykjavik, Iceland, 26–31 May 2014, pages 3222–3228.

Aoife Cahill and Arndt Riester. 2012. Automatically acquiring fine-grained information status distinctions in German. In *Proceedings of the SIGdial 2012 Conference: The 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue,* Seoul, Korea, 5–6 July 2012, pages 232–236.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT’s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP,* Florence, Italy, 28 July–2 August 2019, pages 276–286.

Herbert H. Clark. 1975. Bridging. In *Proceedings of the Conference on Theoretical Issues in Natural Language Processing,* Cambridge, Mass., June 1975, pages 169–174.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Minneapolis, USA, 2–7 June 2019, pages 4171–4186.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307.

M. A. K. Halliday. 1967. Notes on transitivity and theme in English, Part 2. *Journal of Linguistics*, 3:199–244.

Yufang Hou, Katja Markert, and Michael Strube. 2013a. Cascading collective classification for bridging anaphora recognition using a rich linguistic feature set. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* Seattle, Wash., 18–21 October 2013, pages 814–820.

Yufang Hou, Katja Markert, and Michael Strube. 2013b. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Atlanta, Georgia, 9–14 June 2013, pages 907–917.

Yufang Hou, Katja Markert, and Michael Strube. 2014. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* Doha, Qatar, 25–29 October 2014, pages 2082–2093.

Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

Yufang Hou. 2016a. Incremental fine-grained information status classification using attention-based LSTMs. In *Proceedings of the 26th International Conference on Computational Linguistics,* Osaka, Japan, 11–16 December 2016, pages 1880–1890.

Yufang Hou. 2016b. *Unrestricted Bridging Resolution.* Ph.D. thesis, Heidelberg University.

Yufang Hou. 2018a. A deterministic algorithm for bridging anaphora resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* Brussels, Belgium, 31 October– 4 November 2018, pages 1938–1948.

Yufang Hou. 2018b. Enhanced word representations for bridging anaphora resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* New Orleans, Louisiana, 1–6 June 2018, pages 1–7.

Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* Seattle, Wash., 5–10 July 2020, pages 1428–1438.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* Florence, Italy, 28 July–2 August 2019, pages 3651–3657.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel 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),* Hong Kong, China, 3 – 7 November 2018, pages 5803–5808.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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),* Hong Kong, China, 3 – 7 November 2018, pages 4365–4374.

Ivana Kruijff-Korbayová and Mark Steedman. 2003. Discourse and information structure. *Journal of Logic, Language and Information. Special Issue on Discource and Information Structure*, 12(3):149–259.

Knud Lambrecht. 1994. *Information Structure and Sentence Form.* Cambridge, U.K.: Cambridge University Press.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Shared Task of the 15th Conference on Computational Natural Language Learning,* Portland, Oreg., 23–24 June 2011, pages 28–34.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics,* Jeju Island, Korea, 8–14 July 2012, pages 795–804.

Malvina Nissim, Shipara Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the 4th International Conference on Language Resources and Evaluation,* Lisbon, Portugal, 26–28 May 2004, pages 1023–1026.

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,* New Orleans, Louisiana, 1–6 June 2018, pages 2227–2237.

Ellen F. Prince. 1981. Towards a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York, N.Y.

Ellen F. Prince. 1992. The ZPG letter: Subjects, definiteness, and information-status. In W.C. Mann and S.A. Thompson, editors, *Discourse Description. Diverse Linguistic Analyses of a Fund-Raising Text*, pages 295–325. John Benjamins, Amsterdam.

Altaf Rahman and Vincent Ng. 2012. Learning the fine-grained information status of discourse entities. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics,* Avignon, France, 23–27 April 2012, pages 798–807.

Arndt Riester, David Lorenz, and Nina Seemann. 2010. A recursive annotation scheme for referential information status. In *Proceedings of the 7th International Conference on Language Resources and Evaluation,* La Valetta, Malta, 17–23 May 2010, pages 717–722.

Ina Rösiger. 2019. *Computational modelling of coreference and bridging resolution*. Ph.D. thesis, University of Stuttgart.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2011. OntoNotes release 4.0. LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium.