Incremental Fine-grained Information Status Classification Using Attention-based LSTMs

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

Abstract

Information status plays an important role in discourse processing. According to the hearer's common sense knowledge and his comprehension of the preceding text, a discourse entity could be `old`, `mediated` or `new`. In this paper, we propose an attention-based LSTM model to address the problem of fine-grained information status classification in an incremental manner. Our approach resembles how human beings process the task, i.e., decide the information status of the current discourse entity based on its preceding context. Experimental results on the ISNotes corpus (Markert et al., 2012) reveal that (1) despite its moderate result, our model with only word embedding features captures the necessary semantic knowledge needed for the task by a large extent; and (2) when incorporating with additional several simple features, our model achieves the competitive results compared to the state-of-the-art approach (Hou et al., 2013) which heavily depends on lots of hand-crafted semantic features.

1 Introduction

Information status (IS) (Halliday, 1967; Prince, 1981; Nissim et al., 2004) accounts for the familiarity of a discourse entity according to its accessibility to the hearer at a given point in the text, e.g., `old` mentions[1] are known to the hearer and 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; `new` mentions are "not being recoverable from the preceding discourse" (Halliday, 1967).

Information status has attracted a large amount of interests in theoretical linguistics under the framework of *information structure* (Halliday, 1967; Prince, 1981; Prince, 1992; Gundel et al., 1993; Lambrecht, 1994; Birner and Ward, 1998; Kruijff-Korbayová and Steedman, 2003). Many NLP tasks can benefit from knowing information status of discourse entities. Cahill and Riester (2009) improve the performance of generation ranking in German by incorporating features modeling IS. Rahman and Ng (2011) show that a coreference system can profit from IS classification. Baumann and Riester (2013) conduct an empirical study of information status in spoken German and demonstrate that IS can influence prosody in read speech. Hou et al. (2013) model bridging anaphora recognition as a subtask of learning fine-grained information status.

In this paper, we focus on classifying IS on written text because many applications which can benefit from IS concentrate on written texts. We follow the IS scheme for written text proposed by Markert et al. (2012). It adopts the three major IS categories (`old`, `new` and `mediated`) from Nissim et al. (2004) and distinguishes six subcategories for `mediated`. Section 2 provides a brief description of the scheme.

We address the task of fine-grained IS classification via an attention-based LSTM model in an incremental manner. The model resembles human beings' cognitive process of determining IS for discourse entities, i.e., during the process of reading a text from left to right, assign IS for each discourse entity according to its own property and its preceding context.

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

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1880–1890, Osaka, Japan, December 11-17 2016.

Previous approaches on fine-grained IS classification (Markert et al., 2012; Hou et al., 2013) explore world knowledge by integrating hand-crafted semantic features extracted from manually and automatically constructed knowledge bases. One goal of this paper is to investigate in which extent word embeddings learned from large corpora can replace such hand-crafted semantic features. Experimental results on the ISNotes corpus (Markert et al., 2012) show that our model with only word embedding features achieves reasonable results for several IS categories. We further demonstrate that when incorporating with several additional simple features, our model achieves competitive results compared to the state-of-the-art approach (Hou et al., 2013) which heavily depends on lots of hand-crafted semantic features.

2 An Overview of Information Status in ISNotes

ISNotes (Markert et al., 2012) contains 10,980 mentions annotated for information status in 50 texts taken from the Wall Street Journal portion of the OntoNotes corpus (Weischedel et al., 2011). Below we briefly illustrate the definitions of eight IS categories with examples.

A mention is `old` if it is either coreferent with an already introduced entity, or if it is a generic or deictic pronoun.

`Mediated` mentions have not been mentioned before but are not autonomous, i.e., they can only be correctly interpreted by reference to another mention or to prior world knowledge. ISNotes distinguishes six subcategories of `mediated` mentions:

- `mediated/worldKnowledge` mentions are generally known to the hearer. This category includes many proper names, such as *Germany*.
- `mediated/syntactic` mentions are syntactically linked via a possessive relation, a proper name premodification or a PP (prepositional phrase) postmodification to other `old` or `mediated` mentions, such as:

  [[their]$_{old}$ liquor store]$_{mediated/syntactic}$,

  [the [Federal Reserve]$_{mediated}$ boss]$_{mediated/syntactic}$, and

  [the main artery into [San Francisco]$_{mediated}$]$_{mediated/syntactic}$.
- `mediated/bridging` mentions are inferable because a related entity or event (antecedent) has been previously introduced in the discourse, such as **the streets** in Example 1.
- `mediated/comparative` mentions usually include a premodifier that makes clear that this entity is compared to a previous one (antecedent), such as **others** in Example 2.
- `mediated/aggregate` mentions are coordinated mentions where at least one element in the conjunction is `old` or `mediated`, such as *[Not only [George Bush]$_{mediated}$ but also [Barack Obama]$_{mediated}$]$_{mediated/aggregate}$.*
- `mediated/function` mentions refer to a value of a previously explicitly mentioned function (e.g., **3 points** in Example 3). The function needs to be able to rise and fall.

(1) *Oranjemund, the mine headquarters*, is a lonely corporate oasis of 9,000 residents. Jackals roam **the streets** at night ...

(2) As the death toll from last week's temblor climbed to 61, the condition of *freeway survivor Buch Helm*, who spent four days trapped under rubble, improved, hospital officials said. Rescue crews, however, gave up hope that **others** would be found.

(3) IBM shares were down$_{function}$ **3 points**.

`New` mentions are entities that have not yet been introduced in the discourse and that the hearer cannot infer from either previously mentioned entities/events or general world knowledge.

Table 1 shows the IS distribution in ISNotes which contain 1,726 sentences in total.

Mentions	10,980	
old	3237	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 1: IS distribution in ISNotes. The last column indicates the percentage of each IS category relative to the total number of mentions.

3 The Attention-based LSTM Model

In this section we first briefly describe LSTMs in Section 3.1. We then detail our attention-based LSTM model for fine-grained IS classification in Section 3.2.

3.1 LSTMs

Recently, recurrent neural networks (RNNs) with long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997) have been empirically shown to perform well in a range of NLP tasks, such as machine translation (Sutskever et al., 2014), parsing (Vinyals et al., 2015), and sentence compression (Filippova et al., 2015). LSTMs contain special units called memory blocks in the recurrent hidden layer. These memory blocks are designed to avoid vanishing gradients and to remember some long-distance dependencies from the input sequence. The vanilla LSTM model with hidden size k is defined as follows: given a sequence of input $(x_1, ..., x_T)$, LSTMs compute the *h*-sequence and the *m*-sequence using the following equations iteratively from t=1 to T:

$$i_t = sigm(W_1 x_t + W_2 h_{t-1}) \quad (1) \qquad f_t = sigm(W_2 x_t + W_3 h_{t-1}) \quad (2)$$

$$o_t = sigm(W_5 x_t + W_6 h_{t-1}) \quad (3) \qquad i_t\prime = tanh(W_7 x_t + W_8 h_{t-1}) \quad (4)$$

$$m_t = m_{t-1} \odot f_t + i_t \odot i_t\prime \quad (5) \qquad h_t = o_t \odot tanh(m_t) \quad (6)$$

The operator $\odot$ denotes element-wise multiplication, and $sigm$ and $tanh$ are computed element-wise. The matrices $W_1, \ldots, W_8$ and the vector h_0 are the parameters of the model.

3.2 Incremental IS Classification with Attention-based LSTMs

Model. In practice LSTMs still have difficulties to handle long-range dependencies because the model tries to encode the full input sequence into a fixed-length vector. To alleviate this problem, attention-based LSTMs allow the decoder to "attend" the different part of the input sequence when making the prediction. In an IS classification scenario, for each document, the attention-based LSTM model reads the mentions from left to right, and predicts each mention's IS output according to (1) the current mention's state cell and (2) weighted representation of the preceding mentions.

Figure 1 shows the high-level structure of our model. More precisely, for a mention m_i and its preceding t mentions, let two LSTMs read the mentions from left to right. The first LSTM uses the sum of word embeddings as mention representations, whereas the second LSTM uses one-hot vectors as mention representations[2]. Let k_1 and k_2 be the hyper-parameters denoting the size of mention representations and

[2]Another choice is to concatenate the sum of word embeddings with one-hot vectors and use only one LSTM. In practice, we find that encoding them with two LSTMs works better.

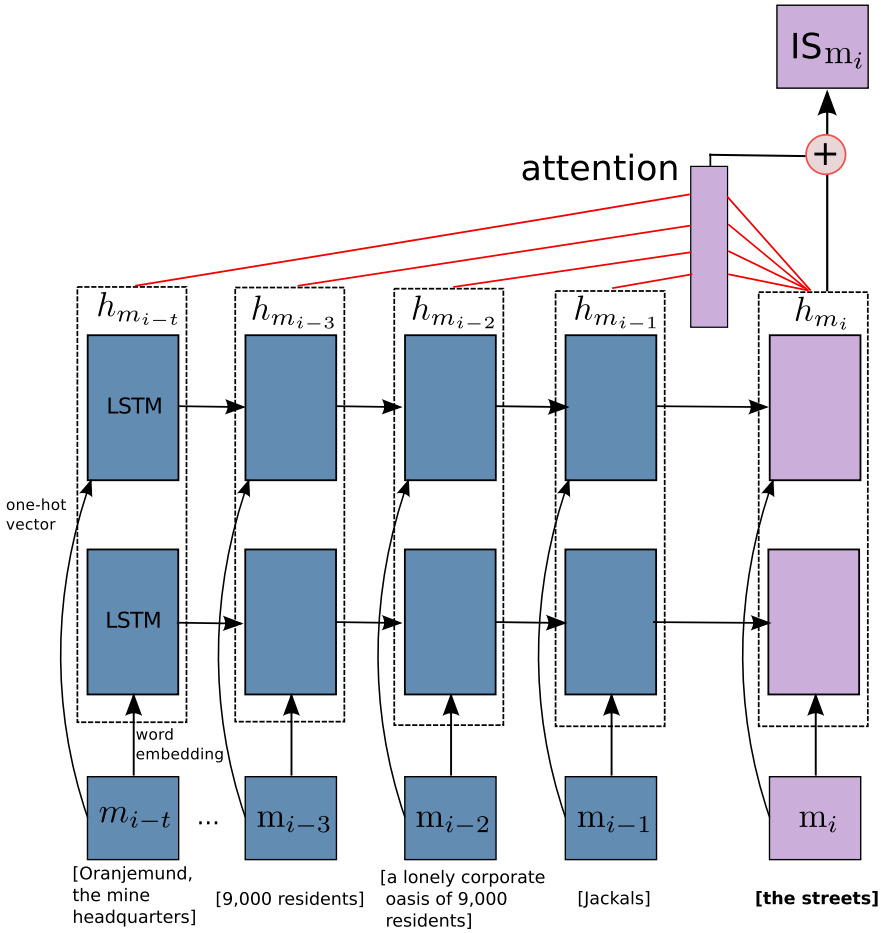

Figure 1: Fine-grained IS classification using attention-based LSTMs.

hidden layers in the two LSTMs respectively[3], and $H_1 \in \mathbb{R}^{k_1 \times (t+1)}$ and $H_2 \in \mathbb{R}^{k_2 \times (t+1)}$ to denote the output vectors from the first LSTM and the second LSTM. We then stack the first t output vectors from the two LSTMs:

$$H = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix}, H \in \mathbb{R}^{(k_1+k_2) \times t} \tag{7}$$

Let $k = k_1 + k_2$, we define an attention vector α over the preceding t mentions and their weighted representation r as follows:

$$M = tanh(W_H H + [(W_{m_i} h_{m_i})_{\times t}]), M \in \mathbb{R}^{k \times t} \tag{8}$$

$$\alpha = softmax(W^T M), \alpha \in \mathbb{R}^t \tag{9}$$

$$r = H\alpha^T, r \in \mathbb{R}^k \tag{10}$$

where h_{m_i} is the stacked output vector of mention m_i from the two LSTMs, the matrices $W_H, W_{m_i} \in \mathbb{R}^{k \times k}$ and the vector $W \in \mathbb{R}^k$ (W^T denotes its transpose) are learned parameters of the model. Note that we repeat the linear transformation of the state cell of mention m_i (i.e., $W_{m_i} h_{m_i}$) t times. As a result, each column in M is the attention representation for each preceding mention $m_j (i - t \leq j < i)$ by combining the output vector h_{m_i} of mention m_i and the output vector of mention m_j (j's column vector in H).

[3]The size of the hidden layer in each LSTM is equal to its mention representation size.

We obtain the final representation of m_i using:

$$\hat{h}_{m_i} = tanh(W_1 r + W_2 h_{m_i}), \hat{h}_{m_i} \in \mathbb{R}^k \quad (11)$$

where the matrices $W_1, W_2 \in \mathbb{R}^{k \times k}$ are the model's parameters. Finally, we use a softmax layer to project $\hat{h}_{m_i}$ into the target space of eight IS classes.

Training instances. For all mentions in a document, we first add a dummy mention with the span of $[-1, -1]$ in the beginning of the document. We then order all mentions according to their end positions in ascending order; if two mentions have the same end position, we order them according to their start positions in descending order. This rule ensures that for embedded mentions, the inside mention is ordered before its parent. Such arrangement of embedded mentions is important because for `mediated/syntactic` and `mediated/aggregate`, a mention's IS label is dependent on the IS labels of its (syntactic) children. Table 2 shows several examples of how embedded mentions are ordered under this rule.

embedded mentions	result of ordering
(1) [[their] liquor store]	[their] – [their liquor store]
(2) [the [Federal Reserve] boss]	[Federal Reserve] – [the Federal Reserve boss]
(3) [the main artery into [San Francisco]]	[San Francisco] – [the main artery into San Francisco]
(4) [[he] and [[his] skilled team]]	[he] – [his] – [his skilled team]–[he and his skilled team]

Table 2: Results of ordering for embedded mentions.

After ordering, we create a training instance for each mention using its preceding mentions as the context. The training instances are created in an incremental manner. For instance, given a document containing the sentence shown in Example 1 (Section 2), the training instances will be (m_0 is the dummy mention):

- m_0 || **[the mine headquarters]**
- m_0*–[the mine headquarters]* || **[Oranjemund, the mine headquarters]**
- m_0*–[the mine headquarters]–[Oranjemund, the mine headquarters]* || **[9,000 residents]**
- m_0*–[the mine headquarters]–[Oranjemund, the mine headquarters]–[9,000 residents]* || **[a lonely corporate oasis of 9,000 residents]**
- m_0*–[the mine headquarters]–[Oranjemund, the mine headquarters]–[9,000 residents]–[a lonely corporate oasis of 9,000 residents]* || **[Jackals]**
- m_0*–[the mine headquarters]–[Oranjemund, the mine headquarters]–[9,000 residents]–[a lonely corporate oasis of 9,000 residents]–[Jackals]* || **[the streets]**
- m_0*–[the mine headquarters]–[Oranjemund, the mine headquarters]–[9,000 residents]–[a lonely corporate oasis of 9,000 residents]–[Jackals]–[the streets]* || **[night]**

Decoding. In the testing stage, given a document and its ordered mentions based on the rule described above, we predict IS classes for these mentions incrementally from left to right. Because a mention's IS could depends on the IS labels of its context mentions, we also encode the IS class as one-hot representation. The gold standard labels and the predicted IS labels of the context mentions are used for training and decoding respectively[4].

[4]The IS class one-hot representation for the target mention is a zero vector during training and decoding.

Network parameters. We use Adam (Kingma and Ba, 2015) for optimization with the learning rate of 0.01. We train all models with 10 epochs using cross-entropy loss. To avoid over-fitting, we apply dropout before and after the LSTM layer with the probability of 0.1. For each mention, we set the maximum number of its context mentions as 50[5]. Therefore we unfold the network 51 times and apply masking for the instances which have less than 50 context mentions.

We use GloVe vectors (Pennington et al., 2014) with 100 dimensions trained on Wikipedia and Gigaword as word embeddings, which we do not optimize during training. Out-of-vocabulary words in the training set and the testing set are set to fixed random vectors. We approximate mention representations fed into the first LSTM by summing embeddings of all words from a mention as Yu and Dredze (2015) show that sum of word embeddings achieves reasonable result to induce phrase embeddings. In ISNotes, around 30% of mentions contain only one word and around 70% of mentions contain less than four words. Mention representations fed into the second LSTM are one-hot vectors of the mentions' features and their IS classes.

4 Experiments

4.1 Experimental Setup

We perform experiments on the ISNotes corpus (Markert et al., 2012). Following Hou et al. (2013), all experiments are performed via 10-fold cross-validation on documents. We use gold standard mentions and the OntoNotes syntactic annotation layer for feature extraction. We report overall accuracy as well as precision, recall and F-measure per IS category. In the following, we describe the baseline and our model with different feature settings.

Baseline. Hou et al. (2013) report the state-of-the-art performance for fine-grained IS classification on ISNotes using collective classification. They explore a wide range of features (34 in total), including a large number of lexico-semantic features as well as a couple of surface features and syntactic features. Hou et al. (2013) observe that bridging anaphors are rarely marked by surface features. Therefore they carefully design discourse structure, lexico-semantic and genericity detection features to capture the phenomenon. The semantic features are extracted from manually or automatically constructed knowledge bases, such as WordNet (Fellbaum, 1998) and the General Inquirer lexicon (Stone et al., 1966).

LSTM. To test how well we can predict IS for mentions without using any hand-crafted features, we only use word embeddings and IS labels in our attention-based LSTM model described in Section 3.2. Specifically, we use mention embeddings (100 dimensions) as the input of the first LSTM. Mention embeddings are obtained by summing word embeddings of all words from a mention, where word embeddings are from GloVe vectors trained on Wikipedia and Gigaword. We use one-hot vectors (8 dimensions) to encode IS labels and use them as the input of the second LSTM.

LSTM+PAR. Hou et al. (2013) show that coordination parent-child relations and other syntactic parent-child relations among mentions are highly effective for `mediated/coordination` and `mediated/syntactic` classes. We use one-hot vectors (2 dimensions) to integrate such parent-child information into the LSTM model described above. Table 3 demonstrates two examples of one-hot representation for parent-child relations.

[[their] liquor store]	$[\text{their}]_{[1,0]} \parallel \textbf{[their liquor store]}_{[1,0]}$
[[he] and [[his] skilled team]]	$[\text{he}]_{[0,1]} - [\text{his}]_{[0,0]} - [\text{his skilled team}]_{[0,1]} \parallel \textbf{[he and his skilled team]}_{[0,1]}$

Table 3: one-hot representations for parent-child relations.

LSTM+PAR+FEAT. We hypothesize that knowledge about a mention's surface and syntactic properties can be useful to decide its information status. Therefore, we add a small feature set (see Table 4) from

[5]In practice, we find that in our model, the results are similar with the maximum number of context mentions as 10, 20, or 50.

Hou et al. (2013) into our attention-based LSTM model (*LSTM+PAR*) using one-hot representations. $f1$-$f5$ are surface and syntactic features, and $f6$-$f8$ are three simple lexical-semantic features. $f6$-$f8$ provide additional semantic knowledge for three mediated classes (i.e., `mediated/comparative`, `mediated/worldKnowledge` and `mediated/function`) that our current mention embedding representations do not capture well.

Feature	**Value**
$f1$ FullPrevMention	{yes, no, NA}
$f2$ PartialPreMention	{yes, no, NA}
$f3$ Determiner	{bare, def, dem, indef, poss, NA}
$f4$ NPtype	{pronoun, common, proper, other}
$f5$ GrammaticalRole	{subject, subjectPassive, pp, other}
$f6$ PreModByCompMarker	{yes, no}
$f7$ IsFrequentProperName	{yes, no}
$f8$ DependOnChangeVerb	{yes, no}

Table 4: A small feature set from Hou et al. (2013).

4.2 Results and Discussion

Results. Table 5 shows the results of our models compared to the baseline. Our model with word embeddings and only a couple of simple features (*LSTM+PAR+FEAT*) performs as good as the state-of-the-art approach (Hou et al., 2013) which explores a wide range of semantic features based on various knowledge resources. It is worth noting that the model with only word embeddings (*LSTM*) achieves an accuracy of 66.8. Also *LSTM* performs similar as the baseline for bridging anaphora recognition under the multi-class classification setting. This indicates that word embeddings in our model do capture certain semantics needed for the task. The improvement in *LSTM+PAR* over *LSTM* confirms the effectiveness of the two parent-child relations for `mediated/syntactic` and `mediated/aggregate` categories.

Comparing the results of *LSTM+PAR+FEAT* to *LSTM+PAR* and *LSTM+PAR+FEAT−wordEmb*, it seems that word embeddings and the small set of simple features (most of them are capturing the surface and syntactic properties of mentions) are complementary. Specifically, `mediated/function` and `mediated/bridging` benefit most from word embeddings which provide useful semantic knowledge to capture these two categories. On the contrary, the feature set (*FEAT*) provides better generalization capability for `old`, `mediated/worldKnowledge` and `mediated/comparative`.

	baseline Hou et al.(2013)			*LSTM*			*LSTM+PAR*			*LSTM+PAR +FEAT*			*LSTM+PAR +FEAT−wordEmb*		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
old	84.4	86.0	**85.2**	75.7	75.6	75.6	77.9	71.2	74.4	85.4	84.9	**85.2**	83.3	85.7	84.5
m/worldKnow.	67.4	77.3	**72.0**	45.6	52.6	48.8	39.8	53.1	45.5	67.1	74.5	70.6	60.4	65.1	62.7
m/syntactic	82.2	81.9	**82.0**	63.6	63.8	63.7	80.4	73.4	76.7	80.8	81.9	81.4	76.4	79.8	78.1
m/aggregate	64.5	79.5	71.2	11.8	35.2	17.7	50.7	65.2	57.1	67.8	84.6	**75.3**	65.9	86.9	74.9
m/function	67.7	72.1	69.8	46.2	57.7	51.3	26.2	53.1	35.1	64.6	76.4	**70.0**	12.3	88.9	21.6
m/comparative	81.8	82.1	**82.0**	15.0	34.9	21.0	14.2	38.7	20.8	77.9	83.1	80.4	78.3	80.8	79.5
m/bridging	19.3	39.0	**25.8**	16.3	36.9	22.6	18.7	34.0	24.1	15.7	32.3	21.1	0.0	0.0	NaN
new	86.5	76.1	**81.0**	80.5	67.3	73.3	76.2	70.8	73.4	87.2	74.8	80.5	85.0	68.2	75.7
acc		**78.9**			66.8			68.6			78.6			75.1	

Table 5: Experimental results of the attention-based LSTM models compared to the baseline. Bolded scores indicate the best performance for each IS class. There is no significant difference between *LSTM+PAR+FEAT* and the baseline at the level of $p < 0.01$ (Statistical significance is measured using McNemar's χ^2 test (McNemar, 1947)).

The incremental prediction mechanism in our model utilizes the (predicted) IS class information of previous mentions when predicting the IS class for the current mention. To gain a better understanding of such mechanism, we conduct an experiment by removing IS label information from our best model (thus *LSTM+PAR+FEAT - ISLabels*). This leads to a mild decrease in the overall accuracy (from

78.6 to 77.7). When looking at the results, we found that the decrease is centered on the categories of `mediated/syntactic`, `mediated/aggregate` and `new`. This confirms that our incremental decoding strategy helps the model to capture the IS label dependencies among mentions better.

Analysis of attention mechanism. We further analyze the attention mechanism in our model *LSTM+PAR* by manually checking some testing examples from one fold. We choose this setting because we want to investigate whether the model can capture long distance relations between mentions without being informed by the features which indicate whether a mention is fully or partially mentioned before.

Figure 2 shows heat maps of several examples that our model predicts correctly[6]. Note that we must take into account that the model only partially relies on representations obtained from attention, i.e., in Equation 11, the final prediction depends on the combination of attention representation as well as the long range contextual representation obtained from LSTM encoders.

It is interesting to see that in the first example, the model attends to several reasonable antecedents for the pronoun "[it]" when predicting its information status. In the second example, when predicting information status for "[the kingdom]", the model focuses on its antecedent "[Saudi Arabia]". In the third and the fourth examples, the model focuses on the syntactic children when predicting information status for "[its percentage share of OPEC production]" and "[Motorola and other companies]".

We also notice that for `old` mentions, when their antecedents do not appear in the preceding context mentions, the weights of attention are more uniformly distributed. Furthermore, for correctly predicted `mediated/comparative` and `mediated/bridging` mentions, we do not observe clear patterns in their attention weights. We assume this is because we have less training data for these two categories. In addition, only a few of them have antecedents occurring in the preceding ten mentions. Therefore, the model seems mainly uses the last output vector (h_{m_i} in Equation 11) for prediction.

5 Related Work

Automatic IS classification. Markert et al. (2012) applied joint inference for IS classification on the ISNotes corpus. Built on this work, Hou et al. (2013) proposed a cascading collective classification algorithm for bridging anaphora recognition with various semantic features. They report the state-of-the-art result for fine-grained IS classification using collective classification.

Rahman and Ng (2012) studied the fine-grained IS classification problem on the Switchboard dialogue corpus (Nissim et al., 2004). They first designed a rule-based system to assign IS classes to mentions. The rule-based system heavily depends on knowledge resources such as FrameNet (Baker et al., 1998), WordNet (Fellbaum, 1998), and ReVerb (Fader et al., 2011). 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).

Our work differs from the above mentioned work in that we explore a new model which resembles human beings' cognitive process for the task and we replace hand-crafted semantic features with word embeddings which were learned from large corpora in an unsupervised manner.

Attention-based RNNs in NLP. Recently, RNNs with attention mechanisms have demonstrated success in various NLP tasks, such as machine translation (Bahdanau et al., 2015), parsing (Vinyals et al., 2015), image captioning (Xu et al., 2015), and textual entailment (Rocktäschel et al., 2016). Attention-based RNNs allow the model to access its internal memory when making predictions. This property is intuitive for our task because in order to decide a discourse entity's information status, we need to access its context and choose one (or none) discourse entity to attend.

[6]Due to the space limitation, we only show plots for maximum number of context mentions at ten. The patterns we observe here are similar for maximum number of context mentions at 20 or 50.

Figure 2: Attention heat maps.

6 Conclusions

We develop an attention-based LSTM model which draws on the recent advances in research on RNNs for fine-grained IS classification. The system imitates how human beings reason information status of a discourse entity based on its preceding context. The results indicate that our model with only pre-trained word embeddings captures semantic knowledge needed for the task by a large extent. Extending the model with several simple features improves the ability of the system, resulting in competitive results on the ISNotes corpus compared to the state-of-the-art approach which explores a broad variety of semantic features.

The model presented here is intuitive for understanding discourse entities. In the future, it would be worthwhile exploring how to extend the system to model other related discourse processing tasks, such as coreference resolution and bridging resolution.

Acknowledgements

The author would like to thank Charles Joachim for fruitful discussions and the anonymous reviewers for their valuable feedback.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR 2015), San Diego, 2015*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics,* Montréal, Québec, Canada, 10–14 August 1998, pages 86–90.

Stefan Baumann and Arndt Riester. 2013. Coreference, lexical givenness and prosody in German. *Lingua*. Accepted.

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

Aoife Cahill and Arndt Riester. 2009. Incorporating information status into generation ranking. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing,* Singapore, 2–7 August 2009, pages 817–825.

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.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* Edinburgh, Scotland, U.K., 27–29 July 2011.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Katja Filippova, Enrique Alfonseca, Carlos Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* Lisbon, Portugal, 17–21 September 2015, pages 360–368.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yufang Hou, Katja Markert, and Michael Strube. 2013. 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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR 2015), San Diego, 2015*.

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.

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.

Quinn McNemar. 1947. Note on the sampling errors of the difference between correlated proportions or percentages. *Phychometrika*, 12(2):153–157.

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.

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,* Doha, Qatar, 25–29 October 2014, pages 1532–1543.

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. 2011. Learning the information status of noun phrases in spoken dialogues. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* Edinburgh, Scotland, U.K., 27–29 July 2011, pages 1069–1080.

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.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phi Blunsom. 2016. Reasoning about entailment with neural attention. In *Proceedings of the 4th International Conference on Learning representations (ICLR 2016),* Caribe Hilton, San Juan, Puerto Rico, 2–4 May 2016.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and Cambridge Computer Associates. 1966. *General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, Mass.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 3104–3112. Curran Associates, Inc.

Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 2773–2781. Curran Associates, Inc.

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.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32th International Conference on Machine Learning,* Lille, France, 6–11 July 2015, pages 2048–2057.

Mo Yu and Mark Dredze. 2015. Learning composition models for phrase embeddings. *TACL*, 3:227–242.