SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue
15-17 August 2017, Saarbrücken, Germany

Dialogue Act Semantic Representation and Classification Using Recurrent Neural Networks

Pinelopi Papalampidi **Elias Iosif** **Alexandros Potamianos**
School of E.C.E., National Technical University of Athens, 15773 Athens, Greece
{el12003, iosife, potam}@central.ntua.gr

Abstract

In this work, we present a model that incorporates Dialogue Act (DA) semantics in the framework of Recurrent Neural Networks (RNNs) for DA classification. Specifically, we propose a novel scheme for automatically encoding DA semantics via the extraction of salient keywords that are representative of the DA tags. The proposed model is applied to the Switchboard corpus and achieves 1.7% (absolute) improvement in classification accuracy with respect to the baseline model. We demonstrate that the addition of discourse-level features enhances the DA classification as well as makes the algorithm more robust: the proposed model does not require the preprocessing of dialogue transcriptions.

1 Introduction

Dialogue Act (DA) classification constitutes a major processing step in Spoken Dialogue Systems (SDS) assisting the understanding of user input. Typically, this is implemented as the assignment of tags to user utterances that (lexically) describe the respective acts. DAs can be regarded as the minimal units of linguistic communication that are directly connected with the speaker's communicative intentions (Searle, 1969). The output of DA classification can be exploited by other SDS components including the modules of natural language understanding and dialogue management.

Various approaches have been used for DA classification including Bayesian Networks (BN), Hidden Markov Models (HMM) (Stolcke et al., 2000), feed-forward Neural Networks (Ji et al., 2016), Decision Trees (Ang et al., 2005) and Support Vector Machines (SVM) (Fernandez and Picard, 2002). The majority of these approaches examined both the utterance meaning as well as the sequence of the utterances within the dialogue. Recently, Deep Neural Networks (DNNs) have been utilized for dialogue act classification (Kalchbrenner and Blunsom, 2013; Lee and Dernoncourt, 2016; Khanpour et al., 2016; Ji et al., 2016) providing a significant increase in classification accuracy in task-independent conversations.

A challenge in the area of DA classification is the construction of models that are domain-agnostic and perform well across different granularities (coarse- vs. fine-grained) of DA tags. In recent deep learning approaches (e.g., (Kalchbrenner and Blunsom, 2013; Khanpour et al., 2016; Lee and Dernoncourt, 2016)) DNNs rely on word embeddings that are generic or randomly set, ignoring domain-specific semantics. In (Lee and Dernoncourt, 2016), the performance of DA systems using various domain generic word embedding schemes was investigated and it was shown that performance depends on the granularity of DA tags.

In this work, we address the incorporation of DA-specific semantics in the framework of RNNs. Specifically, we propose a novel scheme for the automatic encoding of DA semantics via the extraction of a set of semantically salient keywords. Those keywords can be regarded as members of semantic subspaces that correspond to the respective DA. The importance of such keywords being relative to each DA is estimated by a regression model that exploits word embeddings. The classification of an unknown utterance relies on the computation of semantic similarity scores between the utterance words and the aforementioned DA subspaces, which are given as features in the used DNN in addition to typical word embeddings.

The rest paper is organized as follows. In Section 2, the prior work is presented. In Section 3,

10.21437/SemDial.2017-9

both the baseline model (Khanpour et al., 2016; Lee and Dernoncourt, 2016) and the proposed model are described. In Section 4, the experimental dataset as well as the used DA tags are presented. The experimental setup and the related parameters are provided in Section 5, while the evaluation results are presented in Section 6. Section 7 concludes this work.

2 Related Work

The early approaches of DA classification took advantage of lexical information, syntax, semantics, prosody, and dialogue history with manual extraction of the features (Qadir and Riloff, 2011; Stolcke et al., 2000; Jurafsky et al., 1997b; Klaus et al., 1997; Kim et al., 2010; Novielli and Strapparava, 2013). Qadir and Riloff (2011) built speech act classifiers in message board posts utilizing lexical, syntactic and semantic features by creating fixed, topic specific lexicons with keywords. Stolcke et al. (2000) exploited lexical, collocational and prosodic cues, extracted from dialogues, in combination with discourse information of the DA sequence. The reported model is a Hidden Markov Model (HMM), where each HMM state corresponds to a sequential DA, achieving classification accuracy of 71.0% when applied to the Switchboard-DAMSL corpus (Jurafsky et al., 1997a). Novielli and Strapparava (2013) examined the role of affective analysis through affective lexicons in the recognition of DAs. In terms of affective text analysis, semantic features have been extracted based on the distributional semantic models built by Malandrakis et al. (2013).

Recently, the evolution of deep learning allowed the implementation of different models of DNNs in NLP, including the dialogue act classification. Kalchbrenner and Blunsom (2013) used a mixture of Convolutional Neural Networks (CNNs) as a sentence model for the extraction of features from each utterance and Recurrent Neural Networks (RNNs) as a discourse model for the extraction of information about the sequence of the DA. This work improved the state-of-the-art DA classification on Switchboard-DAMSL corpus, reaching 73.9% accuracy. Lee and Dernoncourt (2016) built a model based on RNN and CNN that incorporates the preceding utterances via a two-layer feedforward Artificial Neural Network (ANN) for the extraction of discourse information. Ji et al. (2016) proposed a hybrid architecture that combines an RNN sentence model with discourse information about the relation between two sequential utterances in the form of a latent variable. When the likelihood of the discourse relations derived from the model is maximized, treating the sentence model as a collateral factor in DA classification, an accuracy of 77.0% is achieved. Khanpour et al. (2016) employed a deep Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) structure with pre-trained word embeddings, and reported a classification accuracy of 80.1% outperforming the state-of-the-art.

For testing the various models suggested for DA classification accuracy, a variety of annotation schemes as well as datasets have been utilized (Jurafsky et al., 1997a; Ang et al., 2005; Kim et al., 2015; Henderson et al., 2014). Jurafsky et al. (1997a) provided a dataset annotated with 42 DA tags according to the Dialog Act Markup in Several Layers (DAMSL) (Allen and Core, 1997) annotation scheme. Ang et al. (2005) proposed an annotation scheme of five classes based on the MRDA corpus. However, efforts are made in order to develop a DA annotation scheme that is task-independent and can be used by automatic annotation methods (Bunt et al., 2012; Bunt et al., 2010; Bunt et al., 2017). Nevertheless, there are still limited data annotated based on the principles of these schemes, such as ISO standard 24617-2 and DIT++ (Bunt et al., 2012; Bunt et al., 2010).

3 Proposed Model

The two parts that constitute the proposed model are depicted in Figure 1. The first part (sentence model) creates a vector representation of the utterance based on the LSTM structure suggested by Lei et al. (2015a) and also used by Khanpour et al. (2016). The sentence model uses word embeddings for the similarity computation between the constituent words of utterances and DA tags. This model is detailed in Section 3.1. The second part is a discourse model that classifies the current utterance based on its representation as well as the representations of the preceding ones as proposed by Lee and Dernoncourt (2016). The discourse model is detailed in Section 3.2. To the baseline model we add the semantic representation of the DA tags.

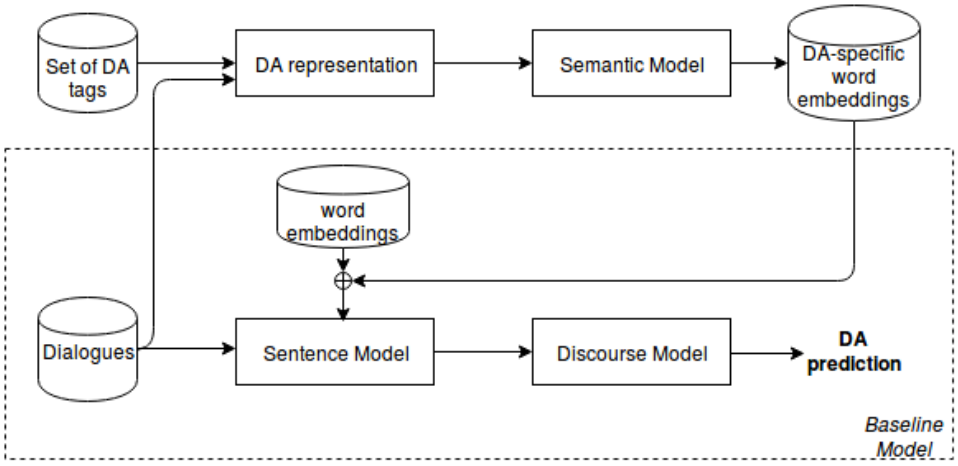

Figure 1: Overview of the proposed model.

3.1 Sentence Model

The proposed sentence model is an extension of the baseline sentence model with DA-specific semantic features as illustrated in Figure 1. The baseline sentence model and the proposed approach of semantic features extraction are described next.

Baseline Sentence Model

The baseline sentence model is depicted in Figure 2. Given an utterance that contains l words, the model converts it into a sequence of l d-dimensional word vectors $X_1, X_2, ..., X_l$. This sequence is given as input to the LSTM network that produces a m-dimensional vector representation s of the utterance. LSTM is a variant of RNN that has the benefit of preserving long-distance dependencies between words and distilling unimportant words from the cell gate through its forget gate layer. In particular, given a sequence $X_1, X_2, ..., X_t, ..., X_l$ of word vectors, for the t^{th} word vector X_t, with inputs h_{t-1} and c_{t-1}, h_t and c_t are computed as follows (Hochreiter and Schmidhuber, 1997):

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \quad (1)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \quad (2)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \quad (3)$$

$$u_t = tanh(W_u x_t + U_u h_{t-1} + b_u), \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot u_t, \quad (5)$$

$$h_t = o_t \odot tanh(c_t), \quad (6)$$

where $W_j \in \Re^{d \times d}, U_j \in \Re^{d \times m}$ for $j \in \{i, f, o, u\}$ are weight matrices, $b_j \in \Re^d$ are bias vectors and $\sigma(\cdot)$ is the element-wise signoid function, $tanh(\cdot)$ is the hyperbolic tangent function and $\odot$ is the element-wise multiplication.

In the pooling layer, all $h_1, h_2, ..., h_t$ vectors that have been computed are combined for the generation of a single vector that represents the utterance. The combination of the h vectors can be produced by applying any of the following schemes: max-pooling, mean-pooling and last-pooling. Max-pooling keeps the element-wise maximum of the h vectors, mean-pooling averages the h vectors and last-pooling keeps the last h vector, namely the h_t vector. In order to obtain longer dependencies between the utterance words, two LSTM cells are stacked as proposed by Graves et al. (2013) and Sutskever et al. (2014). Therefore, the sentence model has two hidden layers.

DA Representation

The typical word embeddings that constitute the input of the sentence model, does not directly model the semantic information about the relation between each utterance word w and each DA tag. Here, we present a semantic model that automatically extracts the domain-specific semantics of w. Specifically, the semantic model computes the semantic similarity between w and each DA. The first step towards calculating semantic similarity between w and each one of the DAs, is the selection of keywords that are representative of the context of the DA tags as described in the following paragraph.

Keyword Selection. In order to automatically

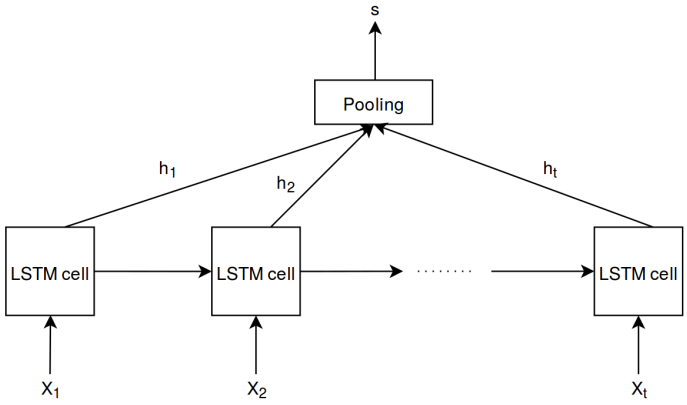

Figure 2: Overview of the baseline sentence model for representing utterance s.

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & d(k_1,w_1)\bar{s}(k_1,t_i) & \cdots & d(k_N,w_1)\bar{s}(k_N,t_i) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & d(k_1,w_K)\bar{s}(k_1,t_i) & \cdots & d(k_N,w_K)\bar{s}(k_N,t_i) \end{bmatrix} \cdot \begin{bmatrix} a_{i0} \\ a_{i1} \\ \vdots \\ a_{iN} \end{bmatrix} = \begin{bmatrix} 1 \\ \bar{s}(w_1,t_i) \\ \vdots \\ \bar{s}(w_K,t_i) \end{bmatrix} \tag{7}$$

determine the keywords that are representative of the DAs, we use the following measurements:

1. Saliency of w, that measures the information content of w in respect to a specific task (DA in this case), as proposed by Gorin (1996):

$$L(w) = \sum_{i=1}^{T} p(t_i|w) log \frac{p(t_i|w)}{p(t_i)} \ , \tag{8}$$

   where $L(w)$ is the saliency of w, T is the number of DA tags, $p(t_i|w)$ is the probability of the i^{th} DA t_i given w, and $p(t_i)$ is the probability of the i^{th} DA t_i,

2. Frequency of w, denoted as $f(w)$,

3. maximum probability of a DA tag given w ($\max_{i=1}^{T} p(t_i|w)$), where t_i is the i^{th} DA.

The keyword extraction is then based on thresholds (see Section 5.1) applied to the product of the saliency of w and its frequency ($S(w)f(w)$) and to the maximum probability of a DA given w ($\max_{i=1}^{T} p(t_i|w)$).

Semantic Model. After determining the keywords, the semantic similarity between w and each DA is computed as follows:

$$s(w,t_i) = \sum_{j=1}^{N} a_{ij} \frac{p(t_i|k_j)p(k_j)}{p(t_i)} d(k_j,w) \ , \tag{9}$$

where $s(w,t_i)$ is the semantic similarity between w and the i^{th} DA t_i normalized in range 0 to 1, N is the total number of keywords and a_{ij} are the weights assigned to each keyword k_j for every DA t_i which are computed according to (7) for every $i \in [1,T]$. $p(t_i|k_j)$ is the probability of the i^{th} DA t_i given the keyword k_j, $p(t_i)$ is the probability of the i^{th} DA t_i, $p(k_j)$ is the probability of the keyword k_j, $\frac{p(t_i|k_j)p(k_j)}{p(t_i)} = p(k_j|t_i)$ is the probability of being keyword k_j representative of the i^{th} DA t_i, normalized in the range 0 to 1 and $d(k_j,w)$ is the cosine similarity between the vectors of w and the keyword k_j.
In (7) where the a weights are calculated, K is the size of the dialogue vocabulary and $\bar{s}(w_k,t_i)$ is the estimated semantic similarity between w_k and the i^{th} DA t_i. $\bar{s}(w_k,t_i)$ is computed by applying (9) and setting the a weights equal to 1.

3.2 Discourse Model

The discourse model is depicted in Figure 3. Let s_i be the vector representation of the i^{th} utterance of the dialogue computed from the sentence model. The sequence s_{i-2}, s_{i-1}, s_i is used as input to a two-layer feedforward ANN. The goal of the discourse model is to predict the DA of the i^{th} utterance ($z_i \in \Re^T$). The output of the first layer of the ANN is computed as follows:

$$y_i = tanh(\sum_{d=0}^{2} W_{-d} s_{i-d} + b_1), \tag{10}$$

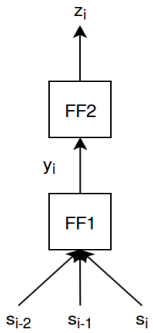

Figure 3: Overview of the discourse model that predicts the DA z_i of utterance s_i.

where $W_0, W_{-1}, W_{-2} \in \Re^{T \times m}$ are the weight matrices, $b_1 \in \Re^T$ is the bias vector, $y_i \in \Re^T$ is the DA representation of the s_i utterance, and T is the number of DAs.

Next, the input of the second layer of the ANN is the vector representation y_i provided by the first layer. The final output of the network is the prediction of the DA for the utterance s_i computed as follows:

$$z_i = softmax(U_0 y_i + b_2), \qquad (11)$$

where $U_0 \in \Re^{T \times T}$ and $b_2 \in \Re^T$ are the weight matrices and bias vector, respectively. For the discourse model, history size of two previous utterances is used for the first layer and no history is taken into account for the second layer as recommended by Lee and Dernoncourt (2016).

4 Experimental Dataset

The dataset used is the Switchboard-DAMSL dataset (Jurafsky et al., 1997a), which is annotated with the 42 DAMSL tags. The Switchboard corpus was originally used for training and testing various speech processing algorithms. Also, it has been used for other tasks such as Automatic Speech Recognition (ASR) (Iyer et al., 1997) and acoustic model adaptation (Povey et al., 2003), including the modeling of DAs (Jurafsky et al., 1997b). This dataset is split into training and test subsets as proposed by Stolcke et al. (2000). The training set comprises of 1,155 dialogues (199,050 utterances) and the test set of 19 dialogues (3,927 utterances) collected over the phone from 500 different speakers. The word-by-word transcriptions are also provided. The topic of discussion between two speakers is introduced by a computer-driven robot agent and the conversation that follows is recorded. About 70 casual topics were introduced. In Table 1, the length of the dialogues (in terms of number of utterances) included in the dataset is presented. A development set was created by randomly selecting 115 dialogues (13,192 utterances) from the training set.

# of Utterances per dialogue	**Train set**	**Test set**
min value	92	187
max value	954	679
mean value	334.6	410.0

Table 1: Switchboard-DAMSL corpus.

In Table 2, representative examples of the eight most frequent DAs are presented. Furthermore, the distribution of the DAs over the dataset is reported in Table 3. As shown in this table, the most frequent DA is the "Statement-non-opinion".

No preprocessing, including tools for stripping the punctuation and changing the capitalization, is applied to the dataset. For the experiments that follow classification accuracy is used as evaluation measurement.

DA tag	**Example**
Statement-non-opinion	There's no one else that works there.
Acknowledge (Backchannel)	Sure.
Statement-opinion	but I think its relevance is pretty limited.
Agree/Accept	That's right.
Abandoned or Turn-Exit	Do you,-
Appreciation	Well good.
Yes-No-Question	So do you have a family too?
Non-verbal	<Laughter>.

Table 2: Examples of the most frequent DAs.

5 Parameter Tuning

In this point, we describe the process for selecting the keywords of the semantic model (see Section 5.1) and the tuning of the hyperparameters of the LSTM baseline model (see Section 5.2). For tuning we used the development set mentioned in Section 4.

DA tag	Train set (%)	Test set (%)
Statement-non-opinion	36.9	31.5
Acknowledge (Backchannel)	18.8	18.2
Statement-opinion	12.7	17.1
Agree/Accept	7.6	8.6
Abandoned or Turn-Exit	5.5	5.0
Appreciation	2.3	2.2
Yes-No-Question	2.3	2.0
Non-verbal	1.7	1.9
Remaining DAs	*12.2*	*13.5*

Table 3: Relative frequency (%) of the DAs.

5.1 Keyword Selection

For the selection of the keywords, classification accuracy is calculated when different thresholds to the metrics described in Section 3.1 are applied. The best performance is achieved when 323 keywords are selected (for $S(w)f(w) = 200$ and $\max_{i=1}^{T} p(t_i|w) = 0.5$). Indicative examples of the selected keywords for the most frequent DAs are presented in Table 4.

DA tag	Selected keywords
Statement-non-opinion	want, can't, work, mine, decided, always, remember
Acknowledge (Backchannel)	huh-uh, huh, yeah, yep, what?, huh?
Statement-opinion	seem, think, scary, ought, worse, difficult
Agree/Accept	true, agree, yes
Abandoned or Turn-Exit	–, -/, -
Appreciation	gosh, dear, wow, kidding
Yes-No-Question	mean?, there?, then?, all?
Non-verbal	<Laughter>, <Noise>, <Clicking>., <sniffing>

Table 4: Examples of automatically selected keywords (shown for most frequent DAs).

5.2 LSTM Parameters

For the implementation of the baseline sentence model (see Section 3.1) the NN packages provided by Lei et al. (2015a) and Lei et al. (2015b) were used. One hyperparameter at a time is tuned while keeping the remaining ones fixed in order to determine the best configuration. Based on findings taken from literature (Khanpour et al., 2016), we initialize the parameters with the following values: word embeddings=200-dimensional vectors with GloVe (Pennington et al., 2014), decay rate=0.7, dropout=0.3, pooling-mechanism=mean-pooling.

Word Embeddings. Keeping the hyperparameters of the LSTM network fixed, different word-to-vector techniques and the dimensionality of the word vectors, that constitute part of the input to the network, are tested. The word vectors are trained either with word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b) method on the GoogleNews corpus or with the GloVe (Pennington et al., 2014) method on the CommonCrawl corpus. Regarding the dimensions of the word embeddings, we use those referred in (Lee and Dernoncourt, 2016) [1]. The word embeddings are then concatenated with the features extracted by the semantic model. The performance for various dimensions is presented in Table 5. As shown in this table, the best performance (75.6%) is achieved when 200-dimensional word embeddings are used. Therefore, for the experiments that follow this setting is used.

Decay Rate. The decay rate is a regularization factor of the update of the network connection weights in order to avoid overfitting of the network. Typically, the decay rate value lies between 0 and 1. In this work, the decay rates that are recommended in the literature (Lee and Dernoncourt, 2016; Khanpour et al., 2016) are examined, as shown in Table 5. The best performance (75.6% accuracy) is achieved with decay rate equal to 0.7 and this setting is used for the rest experiments.

Dropout. For most DNNs, dropout (Hinton et al., 2012) is used as a regularization technique against overfitting. In Table 5, the impact of dropout rate on the classification accuracy is presented for values in the range between 0.0 and 0.5 as proposed in the literature (Lee and Dernoncourt, 2016; Khanpour et al., 2016). The best per-

[1] The word2vec method yields lower classification accuracy (by 0.2%) compared to GloVe and is not reported in Table 5.

Word embeddings	Decay rate	Dropout	Pooling mechanism	Classification Accuracy(%)
50	0.7	0.3	mean	74.7
150				75.4
200				**75.6**
300				75.2
200	0.3	0.3	mean	74.3
	0.5			75.1
	0.9			74.1
200	0.7	0.0	mean	75.4
		0.1		75.4
		0.2		75.5
		0.4		75.4
		0.5		75.2
200	0.7	0.3	max	75.3
			last	75.2

Table 5: Performance of LSTM hyperparameters w.r.t. test set.

formance (75.6% accuracy) is achieved when the dropout rate equals to 0.3 and this setting is used for the experiments that follow.

Pooling mechanism. The various mechanisms that can be used in the pooling layer (max-, mean-, and last-pooling) as described in Section 3.1, are tested. The performance (classification accuracy) for various pooling schemes (max, mean, last) is reported in Table 5. The highest classification accuracy (75.6%) is yielded by the mean-based scheme, which is adopted.

Other Hyperparameters. Here, we briefly mention the settings for a number of other parameters following literature findings (Khanpour et al., 2016). The value of $l2$-regularization is set at $1e-5$ and the $tanh$ function is used for activation in the LSTM cell. Moreover, as reported by Khanpour et al. (2016) changes on the learning rate do not have an impact on the performance of the model. Hence, the learning rate is set at $1e-3$.

6 Evaluation Results

In Table 6, the classification accuracy for both the baseline and proposed model is reported. The highest accuracy (75.6%) is achieved by the proposed model outperforming the baseline by 3.8% when both sentence and discourse information is used. Regarding the sentence-level analysis, the difference between the proposed model and the baseline is even bigger (4.3%). In Table 6 the performance of the baseline model, when applying preprocessing of the dataset, is also presented. In this case, the proposed model still outperforms the baseline by 1.7% accuracy.

Based on the results of Table 6, the proposed model benefits from the additional semantic information. Moreover, it is demonstrated that the proposed model avoids the need for preprocessing of the dataset[2].

The performance of the proposed model is comparable with the state-of-the-art[3] classification accuracy (see Table 7 for an overview) which equals to 77.0% (Ji et al., 2016). An advantage of the present work is the utilization of straightforward feature extraction compared to (Ji et al., 2016) that requires the identification of latent discourse-level features.

7 Conclusions

In this work, we demonstrated the effectiveness of the incorporation of DA-specific semantic features in RNN-based DA classification. Those features were computed with respect to a set of salient keywords meant to semantically represent the DA of interest. The proposed features were found to yield 1.7% (absolute) improvement in classification accuracy with respect to the baseline approach

[2]This was experimentally justified, so, the performance of the proposed model when applying data preprocessing is not reported.

[3]Also, we replicated (use of same model implementation and data) the experiments proposed in (Khanpour et al., 2016) without achieving the same results.

Model	Analysis Level	Preprocessing	Classification Accuracy(%)
Baseline	sentence	✗	69.5
		✓	72.8
Proposed		✗	**73.8**
Baseline	Sentence & discourse	✗	71.8
		✓	73.9
Proposed		✗	**75.6**

Table 6: Performance of the baseline and the proposed model.

Model	Classification Accuracy(%)
Majority classification baseline	*31.6*
Proposed	**75.6**
HMM (Stolcke et al., 2000)	71.0
LSTM (Lee and Dernoncourt, 2016)	69.6
CNN (Lee and Dernoncourt, 2016)	73.1
RCNN (Kalchbrenner and Blunsom, 2013)	73.9
DRLM-joint training (Ji et al., 2016)	74.0
DRLM-conditional training (Ji et al., 2016)	**77.0**
Tf-idf (baseline)	47.3
Inter-annotator agreement	*84.0*

Table 7: Performance of the proposed model and other methods from the literature.

that relies solely on word-level embeddings. Also, we experimentally showed that the discourse-level (specifically, the consideration of current and the previous two utterances) further improves on the baseline performance. Unlike similar approaches presented in the literature, the proposed model does not require any additional tools meant for the preprocessing of dialogues transcriptions.

Regarding future work, we plan to investigate the incorporation of more features derived from deeper discourse analysis. In addition, we aim to further validate the experimental findings of this work by using datasets in languages other than English.

Acknowledgments

This work has been partially supported by the-BabyRobot project supported by the EU Horizon 2020 Programme with grant number 687831.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers.

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the ICASSP*, volume 1, pages I/1061–I/1064.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of LREC*.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of LREC*, pages 430–437.

Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue Act Annotation with the ISO 24617-2 Standard. In *Multimodal Interaction with W3C Standards*, pages 109–135.

Raul Fernandez and Rosalind W Picard. 2002. Dialog act classification from prosodic features using support vector machines. In *Speech Prosody 2002, International Conference*.

Allen L Gorin. 1996. Processing of semantic information in fluently spoken language. In *Proceedings of ICSLP*, volume 2, pages 1001–1004.

Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 273–278.

Matthew Henderson, Blaise Thomson, and Jason Williams. 2014. The second dialog state tracking challenge. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, volume 263.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

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

Rukmini Iyer, Mari Ostendorf, and Herbert Gish. 1997. Using out-of-domain data to improve in-domain language models. *IEEE Signal processing letters*, 4(8):221–223.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. *arXiv preprint arXiv:1603.01913*.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997a. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*, pages 97–102.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997b. Automatic detection of discourse structure for speech recognition and understanding. *1997 IEEE Workshop on Speech Recognition and Understanding*, pages 88–95.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584*.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. *Proceedings of COLING*, pages 2012–2021.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 862–871.

Seokhwan Kim, DHaro Luis Fernando, Rafael E Banchs, Jason Williams, Matthew Henderson, and Koichiro Yoshino. 2015. Dialog State Tracking Challenge 4: Handbook.

Ries Klaus, Coccaro Noah, Shriberg Elizabeth, Bates Rebecca, Jurafsky Daniel, Taylor Paul, Martin Rachel, Van Ess-Dykema Carol, Van Ess-Dykema Carol, and Meteer Marie. 1997. Automatic detection of discourse structure for speech recognition and understanding. *1997 IEEE Workshop on Speech Recognition and Understanding*, pages 88–95.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2015a. Molding cnns for text: non-linear, non-consecutive convolutions. *arXiv preprint arXiv:1508.04112*.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi Jaakkola, Katerina Tymoshenko, Alessandro Moschitti, and Lluis Marquez. 2015b. Semi-supervised question retrieval with gated convolutions. *arXiv preprint arXiv:1512.05726*.

Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. 2013. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nicole Novielli and Carlo Strapparava. 2013. The role of affect analysis in dialogue act identification. *IEEE Transactions on Affective Computing*, 4(4):439–451.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*, volume 14, pages 1532–1543.

Daniel Povey, Philip C Woodland, and Mark JF Gales. 2003. Discriminative MAP for acoustic model adaptation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP03).*, volume 1, pages I–I.

Ashequl Qadir and Ellen Riloff. 2011. Classifying sentences as speech acts in message board posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 748–758.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.