Probabilistic Word Association for Dialogue Act Classification with Recurrent Neural Networks

Nathan Duran[0000−0001−6084−4406] and Steve Battle[0000−0002−7154−7869]

University of the West of England,
Coldharbour Ln, Bristol BS16 1QY
{nathan.duran,steve.battle}@uwe.ac.uk

Abstract. The identification of Dialogue Acts (DA) is an important aspect in determining the meaning of an utterance for many applications that require natural language understanding, and recent work using recurrent neural networks (RNN) has shown promising results when applied to the DA classification problem. This work presents a novel probabilistic method of utterance representation and describes a RNN sentence model for out-of-context DA Classification. The utterance representations are generated from keywords selected for their frequency association with certain DAs. The proposed probabilistic representations are applied to the Switchboard DA corpus and performance is compared with pre-trained word embeddings using the same baseline RNN model. The results indicate that the probabilistic method achieves 75.48% overall accuracy and an improvement over the word embedding representations of 1.8%. This demonstrates the potential utility of using statistical utterance representations, that are able to capture word-DA relationships, for the purpose of DA classification.

Keywords: Dialogue Acts · Neural Networks · Probabilistic

1 Introduction

The notion of a Dialogue Act (DA) originated from John Austin's 'illocutionary act' theory [1] and was later developed by John Searle [25], as a method of defining the semantic content and communicative function of a single utterance of dialogue. The utility of DAs as a set of labels for a semantic interpretation of a given utterance has led to their use in many applications requiring Natural Language Understanding (NLU). In dialogue management systems they have been used as a representation of user and system dialogue turns [5] or as a set of possible system actions [3]. For spoken language translation Kumar et al., [13] utilized the contextual information provided by DAs to improve accuracy in phrase based statistical speech translation.

To facilitate their use in such systems, utterances must first be assigned a single DA label, sometimes called short text classification. Previously, many different approaches have been applied to the DA classification problem, including Support Vector Machines (SVM) and Hidden Markov Models (HMM) [27], n-grams [17] and Bayesian networks [11]. More recently Artificial Neural Network

(ANN) based approaches have led to increased performance, particularly Convolutional Neural Networks (CNN) [10] and Recurrent Neural Networks (RNN) [8, 14, 22]. Many of these approaches consider the DA classification task on both a sentential and discourse level. The sentence level is concerned with how the order and meaning of words compose to form the meaning of a sentence [15]. Similarly, on a discourse level, the order and meaning of sentences compose to form the meaning of sequences in dialogue [24]. Certainly, a given utterance and its associated DA is often directly influenced by the nature of the preceding utterances and the current context of the dialogue. For example, ‘okay’ could be an acknowledgement of understanding or an agreement to a request, the intention is dependent on the utterance it is responding to. This view has led much of the neural network based research to model the semantic content of a sentence, in conjunction with some other contextual information, such as previous utterance or DA sequences, or a change in speaker turn, to predict the appropriate DA for the current utterance [10, 14]. Including such historical and contextual information has been shown to improve classification accuracy [16], and likely must be considered for any sophisticated classification model. However, motivated to examine the importance of different lexical and syntactic features, and their contribution towards DA classification, in this work each utterance is considered individually and out-of-context, that is, without any other contextual or historical information. Further, Cerisara et al., [2] determined that the use of traditional word embeddings, such as GloVe [23] and Word2vec [20], have a limited impact on the DA classification task and therefore, this work explores an alternative approach for utterance representations.

This work presents a simple, yet effective, probabilistic method of utterance representation and DA classification using RNN architectures. The utterance representations are generated from the probability distribution over all DAs for each word in the utterance. Intuitively, each word is represented by a vector of the probabilities that it is associated with each DA, and an utterance is then a matrix formed from the vector representations of the words it contains. This representation method was inspired by the intuition that certain keywords may be associated with certain DA types and therefore act as indicators for the DA of the utterances that contain them [26]. A Long-Short Term Memory (LSTM) network based model is then used to classify the utterances according to their associated DA types. A description of the model and utterance representations can be found in section 3. Experimentally this method is applied to the DA classification task using the Switchboard Dialogue Act (SwDA) corpus (section 4) and yields results comparable with more sophisticated approaches that also consider utterance or dialogue context information. Performance for representations generated from different word frequencies is compared to traditional word vector representations using the same LSTM model in section 5. The following section is a discussion of neural network architectures for DA classification and cue-phrase and n-gram approaches with similar motivations to this probabilistic word representation method.

2 Related Work

The ability of RNN to model long term dependencies in sequential data has led to their widespread use in many Natural Language Processing (NLP) tasks [21], and recently both LSTM and CNN have been applied to the DA classification problem with great success. These approaches commonly employ a combination of a sentence model and a discourse or context model. The sentence model acts at the utterance level and encodes sentences, often from word embeddings, into a fixed length vector representation. The encoded utterances are then used as input to a discourse level model that incorporates some other contextual information and classifies the current utterance. For example, Lee and Dernoncourt [14] experiment with both LSTM and CNN sentence models for encoding short text representations, followed by various ANN architectures for classifying the encoded utterances based on the current, and up to two of the previous, utterances. To try and capture interactions between speakers, Kalchbrenner and Blunsom [10] used a Hierarchical CNN sentence model in conjunction with a RNN discourse model and condition the recurrent and output weights on the current speaker. Liu et al., [16] examine several different CNN and LSTM based architectures that incorporate different context information such as speaker change and dialogue history. Their work shows that including context information consistently yielded improvements over their baseline system.

These approaches all use pre-trained word embeddings to construct utterance representations for the input to a sentence model. While word embeddings carry some semantic similarity information useful for many language classification tasks [18], they do not convey any relational information between the words in an utterance and its associated DA. The work of Cerisara et al., [2] showed that pre-trained embeddings did not help the DA classification task and this is likely due to the word vector training corpora commonly being non-conversational. In an effort to incorporate word-DA relationships into a similar model as those already discussed, Papalampidi, Iosif and Potamianos [22] explored the use of keywords that are representative of DAs. First a set of keywords was constructed based on word frequency and saliency with respect to each DA. The keywords were then used to add a weighting value to pre-trained word embeddings for input to an LSTM sentence model. A two-layer ANN then classified utterances based on the current and preceding two utterances in a similar fashion as [14].

The notion that certain words or phrases can act as indicators for utterance DA labels is more often explored via probabilistic methods. Garner et al., [4] described a theory of word frequencies for dialogue move recognition and concluded that better performance could be achieved using a more involved n-gram model, such as that applied by Louwerse and Crossley [17]. Webb and Hepple [28], selected a set of cue phrases based on the probability of an n-gram occurring within a given DA and keeping only those with the ‘predictivity’ values over a certain threshold. Utterance classification is then performed by identifying the cue phrases it contains and assigning a DA label based on the phrase with the highest predictivity value. The probabilistic methodology in this paper combines aspects of the previously described approaches, sing an LSTM sentence model with utterance representations generated from probabilistic word-DA relations.

3 Model

3.1 LSTM Sentence Model

The sentence model is similar to those used by Papalampidi, Iosif and Potamianos [22], and also Khanpour, Guntakandla and Nielsen [12], and is based on a standard LSTM network as described by Hochreiter and Schmidhuber [7]. A given utterance that contains n words, is converted into a sequence of $n \times m$ dimensional vectors $V_1, V_2, \ldots, V_n$. Where m is either the dimension of the word embeddings, or the number of DA in the case of the probabilistic representations (see section 3.2). The lexical order of the words in the original sentence are represented as a sequence of successive time-steps in V. This sequence is given as input to the LSTM which produces an h dimensional vector at each time-step n, where h is the size of the LSTM hidden dimension. A pooling layer then combines the output vectors from each time-step $h_1, h_2, \ldots, h_n$ into a single vector representation s of the utterance. Finally, a single feed forward layer computes the probability distribution over all DA using the softmax activation function.

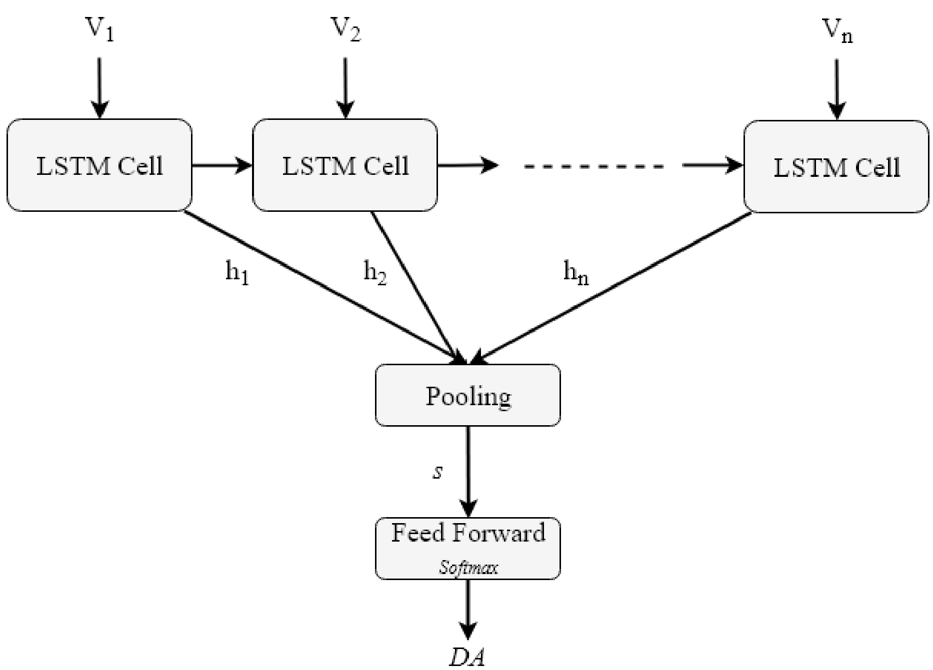

Fig. 1. LSTM Sentence Model

Following the initialisation parameters proposed by Ji, Haffari and Eisenstein [8], with the exception of the bias ($\mathbf{b}$) and output ($\boldsymbol{U}$) matrices, all LSTM weights are initialised in the range $\pm\sqrt{6 \div (d_1 + d_2)}$ where d_1 and d_2 are the input and output dimensions respectively. $\boldsymbol{U}$ is initialised with random uniform distribution in the range ± 0.05 and $\mathbf{b}$ is initialised to 0. Optimisation is performed using the RMSProp algorithm with an initial learning rate $\alpha = 0.001$ and decay rate $\gamma = 0.001$.

3.2 Probabilistic Word Representations

The probabilistic word vector representations are simply generated by calculating a probability distribution over each of the DAs for each word that occurs at, or above, a certain frequency in the corpus vocabulary. First a set of n keywords is created keeping only those that occur at a frequency equal to or greater than a frequency threshold value (see section 5.2). Using this set of keywords, a probability matrix $\boldsymbol{X}$ is created of size $n \times m$, where m is the number of DA used in the corpus. Thus, an element x_{ij} in $\boldsymbol{X}$ will be the probability that the i^{th} word in n appears in an utterance that has the corresponding j^{th} DA tag in m. Each row in $\boldsymbol{X}$ is a probability distribution for the i^{th} word in n for all DAs, and is then effectively a Probabilistic word embedding for creating the utterance representations as described in section 3.1.

4 Experimental Dataset

The Switchboard Dialogue Act (SwDA) contains 1155 transcripts of $\sim$ 5-minute telephone conversations between two speakers that did not know each other and were provided a topic for discussion. Each of the $\sim 205,000$ utterances is annotated with one of 42 DA tags using the Discourse Annotation and Markup System of Labelling (DAMSL) [9]. The transcripts were split into the same 1115 for training and 19 for testing as used by Stolcke et al., [26] and others [8, 10]. The remaining 21 transcripts were used as a validation set and 300 were randomly selected from the training set for development purposes.

Table 1. Datasets

Dataset	# of Transcripts	# of Utterances
Training	1115	192,768
Development	300	51,611
Test	19	4,088
Validation	21	3,196

The transcripts were pre-processed to remove the disfluency (breaks or irregularities), and other annotation symbols, in order to convert each utterance into a plain text sentence. Additionally, any utterances tagged as 'Non-Verbal', such as laughter or coughing, were removed from the transcript, as these do not contain any relevant lexical information. In this work therefore, 41 of the original 42 DAMSL tags were used, reducing the utterance count by $\sim 2\%$.

Table 2. Most frequent Switchboard DA tags.

Dialogue Act	Tag	Example	Count	%
Statement-non-opinion	sd	*Me, I'm in the legal department.*	75,138	37%
Acknowledge (Backchannel)	b	*Uh-huh.*	38,233	19%
Statement-opinion	sv	*I think it's great*	26,422	13%
Abandoned or Turn-Exit	%-	*So, -*	15,545	7%
Agree/Accept	aa	*That's exactly it.*	11,123	5%
Appreciation	ba	*I can imagine.*	4,759	2%
Yes-No-Question	qy	*Do you have special training?*	4,726	2%
Yes answers	ny	*Yes.*	3,031	1%

5 Results and Discussion

5.1 Parameter Tuning

As previously stated, the same LSTM model is used for comparison between the traditional word embedding utterance representations and the Probabilistic representations. Parameters were first tuned using traditional word embeddings and then applied to the Probabilistic representations. Keeping the hyperparameters fixed, different word-to-vector techniques and dimensions are tested. Each parameter is then tuned one at a time to determine the optimum configuration. Both l2-regularization and decay rate were also tested, though were shown to only have a negative impact on performance, and therefore l2-regularization is not used and decay is fixed at $\gamma = 0.001$. For all parameter tuning and utterance representation testing the development set described in section 4 is used. Results shown are averaged over 5 runs of 10 epochs.

Word Embeddings Two pre-trained sets of word embeddings were tested, word2vec [19] trained on the Google News corpus and GloVe [23] trained on a Wikipedia corpus. In addition, a second word2vec embedding was trained on the SwDa corpus itself using the Gensim python package.[1] Dimensions in the range 100-300 were tested and hyperparameters were kept fixed with the values; dropout = 0.3, hidden dimension = 64, and max-pooling. Table 3 shows the best classification accuracy was achieved using word2vec trained on Google News with 300 dimensions and this is adopted for the rest of the experiments.

[1] https://radimrehurek.com/gensim/

Table 3. Word embeddings performance.

Word Embeddings	Embedding Dimension	Test Set Accuracy %	Validation Set Accuracy %
GloVe Wiki	100	70.62	74.22
	200	71.31	74.98
	300	70.45	73.68
Word2vec Google News	100	69.41	72.97
	200	71.14	75.06
	300	**71.65**	**75.07**
Word2vec SwDa	100	68.93	72.58
	200	68.71	72.29
	300	69.18	72.81

Dropout Dropout is a regularisation method generally used to prevent overfitting in neural networks [6]. The results in Table 4 concur with others' findings [14, 12, 22] that a value of 0.3 is optimal.

LSTM Hidden Dimension The LSTM hidden state corresponds to the dimensionality of the output vectors of an LSTM cell at each time-step $h_1, h_2, \ldots, h_n$ and therefore determines the dimension of the utterance representation s that is generated by the pooling layer. Table 4 shows minimal impact on performance provided the hidden dimension is close to the maximum sentence length (~ 107 words).

Pooling Two different pooling mechanisms are tested. Max-pooling keeps the element-wise maximum of the h vectors output by the LSTM and mean-pooling averages the h vectors.

Table 4. Hyperparameter tuning performance.

Dropout	Hidden Dimension	Pooling	Test Set Accuracy %	Validation Set Accuracy %
0.0	64	Max	70.91	74.45
0.1			71.36	74.82
0.2			71.58	74.82
0.3			71.62	75.17
0.4			71.26	74.97
0.3	128	Max	**72.32**	**75.66**
	256		72.01	75.51
0.3	128	Mean	68.51	72.09

5.2 Probabilistic Word Embeddings

Different word frequency thresholds (see section 3.2) were tested to explore the performance impact on the probabilistic word representation vectors. The thresholds are simply an indication of the number of times a given word occurs in the SwDa corpus. Table 5 shows accuracy tends to decrease with larger thresholds. This is likely due to data for the utterance representations becoming too sparse as fewer words are included in the probability matrix. For example, a threshold of 2 eliminates around half of the words in the vocabulary. The thresholds minimum value was kept at 2 to maximise the likelihood of words appearing in at least one of the test datasets.

Table 5. Performance for different word frequency thresholds.

Word Frequency	Test Set Accuracy %	Validation Set Accuracy %
2	**74.68**	**77.38**
4	74.14	77.1
6	73.66	75.7
8	73.01	76.35
10	72.69	76.04

Table 6 shows a subset of the probability matrix created using the method described in section 3.2 for the SwDa corpus. It can be seen that certain words correlate significantly with specific DAs. These are particularly useful features for differentiating between DA that are otherwise semantically similar, for example, *Statement-opinion* and *Statement-non-opinion*. However, certain DA's such as *Abandoned/Turn-Exit* do not correlate strongly with any words.

Table 6. Example of word probabilities for the five most common DA's.

	Statement Non-Opinion	Acknowledge (Backchannel)	Statement Opinion	Abandoned Turn-Exit	Agree Accept
My	86.67	0.03	4.88	1.52	0.08
Yeah	0.08	71.49	0.07	1.54	16.68
Should	26.22	0.00	59.19	0.40	0.40
Um	36.94	15.56	9.14	20.30	0.78
True	5.85	0.30	14.67	0.10	62.83

5.3 Results

Evaluation of the probabilistic and pre-trained word embedding representations was performed on the full training set and results shown are an average over 10 runs for the test dataset. Table 7 shows the highest classification accuracies achieved on the test dataset for both word representation methods using the RNN model. The proposed model trained with word embeddings resulted in a similar accuracy (73.68%) as the sentence model in the work of Papalampidi et al., [22]. The model trained on the Bayes representation shows an improvement of 1.8% over the word embeddings baseline. Though direct comparisons are difficult, due to differences in pre-possessing, models and other methodology, Table 7 also shows the probabilistic model is comparable to methods from the literature where context information was also used.

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

Model	Classification Accuracy %
Sentence Level	
Proposed LSTM - Probabilistic	**75.48**
Proposed LSTM - Word Embeddings	73.68
Sentence (Papalampidi et al., 2017)	73.8
Sentence and Discourse Level	
Sentence and Discourse (Papalampidi et al., 2017)	75.6
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**
Inter-annotator Agreement (Stolcke et al., 2000)	84.0
Majority DA baseline	32.2

6 Conclusion

This work has presented a novel probabilistic approach to utterance representation, and an LSTM sentence model, for the purpose of DA classification. When applied to the SwDA corpus in an out-of-context fashion the probabilistic representations improve DA classification accuracy by 1.8% when compared to traditional word embeddings. Further, the overall highest classification accuracy achieved is competitive with approaches from the literature using more sophisticated classifier models that also consider contextual information, and improves on previously published results that only use a sentence level model. These findings also concur with previous work [2] to show that the traditional word embedding approach for utterance representation may not improve accuracy for the DA classification task. This highlights the need to find alternative representation methods for DA classification, such as the proposed probabilistic keyword-DA relationships.

Regarding future work, it would be beneficial to determine whether using probabilistic representations yields similar improvements in accuracy when using additional contextual information, and more sophisticated discourse and sentence models, which have been shown to improve results [14]. Additionally, it may be valuable to determine the portability of the approach by applying keywords gathered from one corpus to DA classification on another distinct corpus [29]. This would help to determine if certain keywords are able to generalise to new corpora, and perhaps reduce the amount of training data required.

References

1. Austin, J.L.: How To Do Things With Words. Oxford University Press, London (1962)
2. Cerisara, C., Král, P., Lenc, L.: On the Effects of Using Word2vec Representations in Neural Networks for Dialogue Act Recognition. Comput. Speech Lang. **47**(July), 175–193 (2017). https://doi.org/10.1016/j.csl.2017.07.009
3. Cuayáhuitl, H., Yu, S., Williamson, A., Carse, J.: Deep Reinforcement Learning for Multi-Domain Dialogue Systems. In: NIPS Work. Deep Reinf. Learn. pp. 1–9. Barcelona (2016)
4. Garner, P.N., Browning, S.R., Moore, R.K., Russell, M.J.: A Theory of Word Frequencies and its Application to Dialogue Move Recognition. In: 4th Int. Conf. Spok. Lang. Process. (ICSLP 96). ISCA, Philadelphia, PA (1996). https://doi.org/10.1109/ICSLP.1996.607999
5. Griol, D., Hurtado, L., Segarra, E., Sanchis, E.: A Statistical Approach to Spoken Dialog Systems Design and Evaluation. Speech Commun. **50**(8-9), 666–682 (2008). https://doi.org/10.1016/j.specom.2008.04.001
6. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv (2012). https://doi.org/arXiv:1207.0580
7. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Comput. **9**(8), 1735–1780 (1997)
8. Ji, Y., Haffari, G., Eisenstein, J.: A Latent Variable Recurrent Neural Network for Discourse Relation Language Models. In: Proc. NAACL-HLT 2016. pp. 332–342. ACL, San Diego (2016)
9. Jurafsky, D., Shriberg, E., Biasca, D.: Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. Tech. rep. (1997)
10. Kalchbrenner, N., Blunsom, P.: Recurrent Convolutional Neural Networks for Discourse Compositionality. In: Proc. Work. Contin. Vector Sp. Model. their Compos. pp. 119–126. ACL, Sofia, Bulgaria (2013). https://doi.org/10.1109/ICCV.2015.221
11. Keizer, S.: A Bayesian Approach to Dialogue Act Classification. In: BI-DIALOG 2001 Proc. 5th Work. Form. Semant. Pragmat. Dialogue. pp. 210–218 (2001)

12. Khanpour, H., Guntakandla, N., Nielsen, R.: Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. In: COLING 2016, 26th Int. Conf. Comput. Linguist. pp. 2012–2021. Osaka (2016)
13. Kumar, V., Sridhar, R., Narayanan, S., Bangalore, S.: Enriching Spoken Language Translation with Dialog Acts. In: Proc. 46th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol. Short Pap. - HLT '08. p. 225. No. June (2008). https://doi.org/10.3115/1557690.1557755
14. Lee, J.Y., Dernoncourt, F.: Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. In: NAACL 2016 (2016)
15. Li, D.: The Pragmatic Construction of Word Meaning in Utterances. J. Chinese Lang. Comput. **18**(3), 121–137 (2012)
16. Liu, Y., Han, K., Tan, Z., Lei, Y.: Using Context Information for Dialog Act Classification in DNN Framework. In: Proc. 2017 Conf. Empir. Methods Nat. Lang. Process. pp. 2160–2168. ACL, Copenhagen, Denmark (2017)
17. Louwerse, M., Crossley, S.: Dialog Act Classification Using N-Gram Algorithms. In: FLAIRS Conf. 2006. pp. 758–763. Melbourne Beach (2006)
18. Mandelbaum, A., Shalev, A.: Word Embeddings and their Use in Sentence Classification Tasks. arXiv (2016)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and Their Compositionality. In: NIPS'13 Proc. 26th Int. Conf. Neural Inf. Process. Syst. pp. 3111–3119. Lake Tahoe (2013). https://doi.org/10.1162/jmlr.2003.3.4-5.951
20. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv (2013). https://doi.org/10.1162/153244303322533223
21. Mikolov, T., Karafiát, M., Burget, L., Khudanpur, S.: Recurrent neural network based language model. In: INTERSPEECH. pp. 1045–1048. Makuhari (2010)
22. Papalampidi, P., Iosif, E., Potamianos, A.: Dialogue Act Semantic Representation and Classification Using Recurrent Neural Networks. In: SEMDIAL 2017 Work. Semant. Pragmat. Dialogue. pp. 77–86. No. August (2017). https://doi.org/10.21437/SemDial.2017-9
23. Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. pp. 1532–1543 (2014). https://doi.org/10.3115/v1/D14-1162
24. Schegloff, E.A.: Sequence Oranization in Interaction: A Primer in Conversation Analysis I. Cambridge University Press, Cambridge (2007). https://doi.org/10.1017/CBO9780511791208
25. Searle, J.: Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, London (1969)
26. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., Meteer, M.: Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. Comput. Linguist. **26**(3), 339–373 (2000). https://doi.org/10.1162/089120100561737
27. Surendran, D., Levow, G.A.: Dialog Act Tagging with Support Vector Machines and Hidden Markov Models. In: Interspeech 2006 9th Int. Conf. Spok. Lang. Process. pp. 1950–1953. Pittsburgh (2006)
28. Webb, N., Hepple, M.: Dialogue Act Classification Based on Intra-Utterance Features. In: Proc. AAAI Work. Spok. Lang. Underst. (2005)
29. Webb, N., Liu, T.: Investigating the portability of corpus-derived cue phrases for dialogue act classification. In: Proc. 22nd Int. Conf. Comput. Linguist. - Vol. 1. pp. 977–984. No. August, Manchester (2008)