INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Multi-lingual Dialogue Act Recognition with Deep Learning Methods

Jiří Martínek[1,2], *Pavel Král*[1,2], *Ladislav Lenc*[1,2], *Christophe Cerisara*[3]

[1]Dept. of Computer Science & Engineering
University of West Bohemia
Plzeň, Czech Republic
[2]NTIS - New Technologies for the Information Society
University of West Bohemia
Plzeň, Czech Republic
[3]LORIA UMR 7503
BP 239 - 54506 Vandoeuvre, France
{jimar,pkral,llenc}@kiv.zcu.cz,cerisara@loria.fr

Abstract

This paper deals with multi-lingual dialogue act (DA) recognition. The proposed approaches are based on deep neural networks and use word2vec embeddings for word representation. Two multi-lingual models are proposed for this task. The first approach uses one general model trained on the embeddings from all available languages. The second method trains the model on a single pivot language and a linear transformation method is used to project other languages onto the pivot language. The popular convolutional neural network and LSTM architectures with different set-ups are used as classifiers. To the best of our knowledge this is the first attempt at multi-lingual DA recognition using neural networks. The multi-lingual models are validated experimentally on two languages from the Verbmobil corpus.

Index Terms: CNN, deep learning, dialogue act, LSTM, multilinguality, word embeddings, word2vec

1. Introduction

Nowadays, the importance of multi-lingual and cross-lingual natural language processing (NLP) methods is still growing. Another important research direction is the usage of deep neural networks that learn parameters implicitly and do not require manual feature engineering. Both research directions respectively help to significantly reduce the amount of human annotation efforts and improve the applicability of the models to various corpora and contexts. Many researchers have proposed multi-lingual approaches based on neural networks for a wide spectrum of NLP tasks, including document classification [1], named entity recognition [2] and semantic role labelling [3]. Unfortunately, to the best of our knowledge, research in the multi-lingual automatic DA recognition field is scarce.

DA recognition is an important step in dialogue understanding and it plays a pivotal role in dialogue management [4]. Any improvement in this task may increase the performance of the whole dialogue system. In this paper we propose and compare several methods for multi-lingual and cross-lingual DA recognition. The methods utilize deep neural networks and word2vec embeddings are used for word representation. The first approach trains one general model on annotated DAs from all available languages. This model is thus able to perform DA recognition in multiple languages simultaneously. The second method trains the model only on one language and cross-linguality is achieved by a linear semantic space transformation.

We employ two standard neural network topologies with different set-ups, namely the convolutional neural network (CNN) and the long short-term memory (LSTM), and we compare and evaluate them on the Verbmobil corpus [5]. Implementations of all presented methods are publicly available for research purposes.

2. Related Work

Traditional DA recognition methods usually create complex handcrafted features using one of, or a combination of the following types of information:

- lexical and syntactic information
- prosodic information
- dialogue history

Several lexical models are proposed including Bayesian approaches such as n-gram language models [6, 7, 8] and also non-Bayesian methods – semantic classification trees [9], transformation-based learning (TBL) [10] or memory-based learning [11]. Syntactic features that are created using a full parse tree are considered for instance in [12]. Prosodic information is often used to provide additional clues to recognize DAs as presented for instance in [13]. Dialogue history (the sequence of DAs) can been used to predict the most probable next dialogue act and it can be modeled for example by hidden Markov models [7] or Bayesian networks [14].

Nowadays, DA recognition is often realized with simple word features and sophisticated neural models including popular CNNs and an LSTM. The features are represented by word embeddings provided by word2vec [15] or glove [16]. Both word2vec and glove are used in [17] where a CNN (or an LSTM) is used to create a vector representing a DA. This vector is then fed into a feed forward network utilized for classification. The experiments on several DA corpora have shown that the CNN performs slightly better than the LSTM in this case. Another approach using only raw word forms as an input is presented in [18]. The LSTM with word2vec embeddings is used to model the DA structure as well as for DA prediction with very good results. A similar LSTM model with word2vec embeddings has been also proposed for DA recognition in [19]. The authors experimented with different embedding sizes and network hyper-parameters. Duran et al. [20] propose new features called ”probabilistic word embeddings” which are based

http://dx.doi.org/10.21437/Interspeech.2019-1691

on word distribution across DA-set. The experiments show that these features perform slightly better than word2vec.

The above mentioned approaches are mainly evaluated on English language using Switchboard (SwDA) [21] or Meeting Recorder Dialog Act (MRDA) [22] corpora. Some methods are tested on Spanish (see DIHANA [23] corpus), Czech [24], French [25] or on German (see Verbmobil [4] corpus) languages.

These corpora are annotated according to different annotation schemes and they contain different DA labels, although relevant efforts towards a standardization of DA annotations have been made [26]. Their direct usage for multi-lingual DA recognition is difficult, because a mapping between the different tagsets is required. To the best of our knowledge, only the Verbmobil corpus contains a large enough number of annotated dialogues in multiple languages. Mapping DA annotations that are both satisfactory from the linguistic point of view and easily usable computationally is challenging and complex, therefore we have thus chosen the robust and well-known Verbmobil corpus in our experiments.

3. Multi-lingual DA Recognition

This section starts by describing the two methods we use to achieve multi-linguality. The neural network architectures are described next.

3.1. Multi-lingual Model

Let $\mathbb{L} = \{L_1, L2, ..., L_M\}$ be a set of languages with available annotated DAs and $\mathbb{T}_{L_i}$ be the set of DAs for language L_i. Pooling together all of these labels into a single set $\mathbb{T} = \bigcup_{i=1}^{M} \mathbb{T}_{L_i}$ enables to train a multi-lingual classifier that assigns to any input text in any language $L_i \in \mathbb{L}$ a single label from $\mathbb{T}$. Such a model is able to recognize DAs in arbitrarily many languages but it is necessary to retrain the model when a new language is added.

3.2. Cross-lingual Model

The cross-lingual model relies on a semantic space transformation. It is indeed possible to transform the lexical semantic space of any language so that word representations of similar concepts in different languages are close. Based on our previous work [27] we chose the canonical correlation analysis (CCA) [28] method. It is a technique for multivariate data analysis and dimensionality reduction, which quantifies the linear associations between a pair of random vectors. It can be used for a transformation of one semantic space to another.

The DA recognition model is trained on a single pivot language. The test examples from any language are then projected into the target pivot language. It thus allows classifying DAs in any language from within the transformed semantic space. Retraining the DA recognition model is not necessary when a new language is considered.

3.3. DA Representation

Word2vec embeddings are used to encode word semantics. For the cross-lingual scenario, we create a vocabulary V of the $|V|$ most frequent words in the pivot language used for training. In the multi-lingual case the vocabulary is shared and consists of the union of the vocabularies of all available languages.

In order to benefit from parallel GPU processing, the input texts have a constant length W. Therefore, utterances that are longer than W words are truncated, while utterances with less words are padded.

The input to each proposed neural network model is either a sequence of W vocabulary indexes, when the word embedding matrix is considered as part of the model's parameters; or directly a sequence of W embedding vectors when these embeddings are considered as constant.

The advantage of the former input is the possibility to fine-tune the word vectors, while the latter option allows us to use the transformed semantic spaces seamlessly.

3.4. Neural Network Topologies

3.4.1. *Convolutional Neural Networks*

We use two CNN networks with different configurations. The first one is the model presented in [27] where it was used for document classification. We have modified the size of the convolutional kernels to adapt them to the dialogue acts domain. In such a domain, we usually work with much shorter inputs so we use a smaller kernel – $(4, 1)$ as shown in Figure 1.

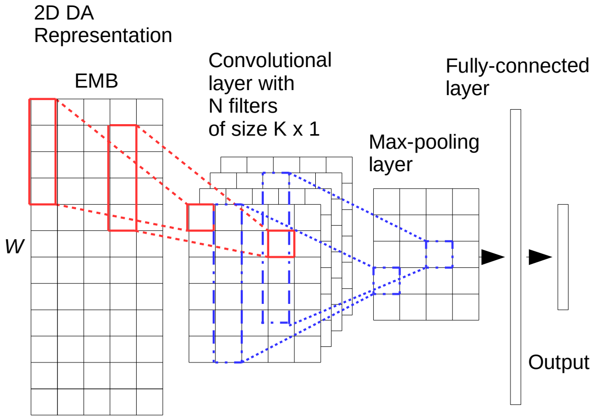

Figure 1: *CNN$_1$ architecture*

We use 40 convolutional kernels with *relu* activation. A final fully-connected layer after the convolutional one consists of 256 neurons, which are further concatenated with the previous vector when we consider the history (the previous DA). We use categorical cross-entropy as a loss function and softmax activation in the output layer. We refer to this architecture as CNN$_1$.

The second configuration follows Kim [29]. It uses three sizes of convolutional kernels – $(3, EMB)$, $(4, EMB)$ and $(5, EMB)$ where EMB is the embedding dimensionality. 100 kernels of each size are computed simultaneously and their outputs are merged and fed into a fully connected layer. The final layer is the same as in the previous case. We refer to this model as CNN$_2$.

3.4.2. *Bidirectional Long Short-Term Memory*

The second approach exploits a Bidirectional LSTM layer. The representation of the input and the embedding layer are the same as for the CNNs. The core of this model is the Bi-LSTM layer with 100 units (i.e. 200 units in total for both directions). The word embedding representation of the input (with 15×300 size) is fed into the Bi-LSTM layer, which outputs a single vector of 200 dimensions. This vector is then concatenated with the predicted dialogue act class of the previous sentence encoded in the form of a one-hot-vector. If the dialogue act has no history (e.g. initial dialogue act), a zero vector is used. The output layer has a softmax activation function. Figure 2 shows this model's architecture. This model is trained using teacher forcing, while decoding is greedy.

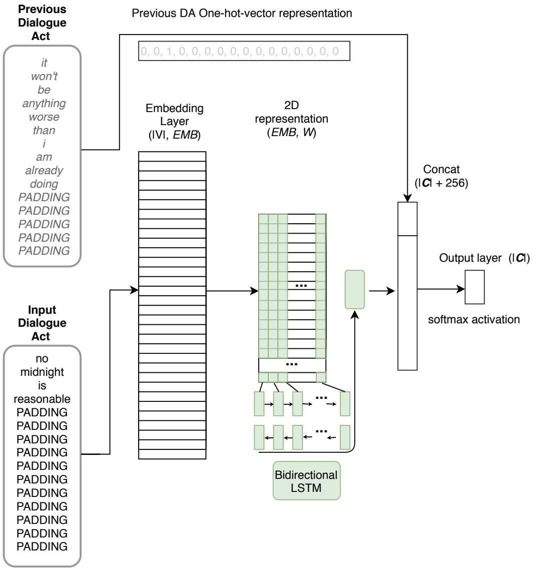

Figure 2: *Bidirectional LSTM architecture*

4. Experiments

4.1. Multi-lingual Verbmobil Corpus

This corpus [5] was created within the Verbmobil project the goal of which was the development of a mobile application for translation of spontaneous dialogues.

It is composed of English, German as well as Japanese dialogues, however our version downloaded from LDC[1] contains only English and German utterances annotated with DA labels. Therefore, we evaluate the proposed approaches on English and German. Statistical information about this corpus is depicted in Table 1.

Table 1: *Corpus statistical information*

	English		German	
unit	Training	Testing	Training	Testing
dialogue #	6 485	940	15 513	622
DA #	9 599	1 420	32 269	1 460
word #	79 506	11 086	297 089	14 819

This dataset is annotated with 42 dialogue acts, which are grouped into the 16 following classes: *feedback*, *greet*, *inform*, *suggest*, *init*, *close*, *request*, *deliberate*, *bye*, *commit*, *thank*, *politeness_formula*, *backchannel*, *introduce*, *defer* and *offer*.

The corpus is very unbalanced. In both languages, there are four dominant DAs (namely *feedback*, *suggest*, *inform*, *request*) which represent almost 80% of the corpus size. Figure 3 shows the DA distribution in the German training part. A similar distribution is obtained on the English corpus.

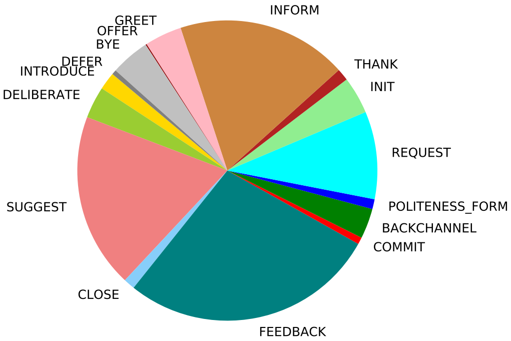

Figure 3: *Distribution of tags in the German training corpus part*

4.2. Experimental Set-up

We use word2vec vectors trained on the English and German Wikipedia to initialize the word embeddings in our models. Sentences are truncated or padded to 15 words in all experiments. The vocabulary size is set to 10,000. The model hyperparameters are fine-tuned on 1,000 of randomly selected utterances from the training data set.

We evaluate all models with and without information from the dialogue history, which consists of the dialogue act that has been predicted at the previous sentence.

Although most related works use the accuracy (Acc) measure, we further compute the F1 score (macro), because the corpus is unbalanced and therefore the F1 score is more relevant. We run all experiments 10 times and the results are averaged.

4.3. Multi-lingual Model Results

This series of experiments shows results of the multi-lingual model presented in Section 3.1. Table 2 reports the performance of the models with static word2vec embeddings while Table 3 presents the results with fine-tuned embeddings. These tables show that, generally, fine-tuning word2vec embeddings does not bring any improvement for DA recognition. The relatively high differences between the accuracy and F1 score are caused by the significant corpus unbalances. Another interesting observation is that the dialogue history helps for DA recognition in all but a few cases and that the best neural classifier is the Bi-LSTM network.

4.4. Cross-lingual Model Results

Table 4 shows the results of the cross-lingual model presented in Section 3.2. The scores of the cross-lingual model are significantly lower than the scores of the multi-lingual methods reported in Tables 2 and 3. The best reported accuracy is obtained by the Bi-LSTM network and it is close to 60% when we use the German part of the corpus for training (pivot language) and the English dataset for testing. Low F1 score occurred because of the poor results of infrequent DAs, which do not impact much the accuracy values. The lower results for the English →

[1] https://www.ldc.upenn.edu/

Table 2: *Multi-lingual DA recognition with static word2vec embeddings*

		CNN$_1$				**CNN$_2$**				**Bi-LSTM**			
		With History		**No History**		**With History**		**No History**		**With History**		**No History**	
Train	Test	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
en	en	72.1	58.9	72.2	58.4	74.5	**65.8**	74.3	65.1	**74.9**	60.5	74.1	59.9
de	de	72.5	**60.8**	71.8	58.2	71.9	57.5	70.8	56.6	**74.3**	59.3	73.6	59.4
en+de	de	72.0	59.9	71.2	57.7	70.9	57.5	71.1	54.6	**74.3**	**61.7**	73.2	60.6
en+de	en	70.3	55.1	70.0	55.5	71.4	57.1	70.7	**58.6**	**72.8**	58.5	72.6	57.4

Table 3: *Multi-lingual DA recognition with fine-tuned word2vec embeddings*

		CNN$_1$				**CNN$_2$**				**Bi-LSTM**			
		With History		**No History**		**With History**		**No History**		**With History**		**No History**	
Train	Test	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
en	en	72.2	**69.2**	72.2	68.4	**73.7**	68.4	72.1	59.1	73.5	67.2	72.7	67.2
de	de	72.7	59.2	71.7	57.7	72.1	59.1	72.6	**60.4**	**74.9**	57.2	74.3	59.1
en+de	de	71.8	**60.8**	70.8	58.9	70.8	58.4	71.7	58.8	**72.7**	58.2	71.4	58.3
en+de	en	**69.2**	61.2	68.6	58.6	69.6	61.2	68.7	60.2	68.5	60.1	**69.2**	**63.1**

Table 4: *Cross-lingual DA recognition based on CCA transformation*

		CNN$_1$				**CNN$_2$**				**Bi-LSTM**			
		With History		**No History**		**With History**		**No History**		**With History**		**No History**	
Train	Test	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
en	de	30.7	11.9	**34.3**	13.9	31.5	14.5	31.2	15.4	34.0	16.4	34.0	**17.0**
de	en	55.1	26.4	54.4	25.5	53.9	28.3	53.0	27.4	**58.6**	**37.1**	57.5	33.7

German direction can be explained by the significantly smaller corpus size for training. This table further shows that dialogue history slightly helps for DA recognition and that the Bi-LSTM significantly outperforms the two other CNN models.

5. Comparison with Related Work

Table 5 compares the performances of the proposed models with several state-of-the-art systems.

Table 5: *Comparison with the state of the art [accuracy in %].*

Method	Acc %
n-grams + complex features [6]	74.7
TBL + complex features [10]	71.2
ME + BoW features	49.1
LSTM + w2vec features [18]	74.0
CNN + w2vec features (proposed)	74.5
Bi-LSTM + w2vec features (proposed)	**74.9**

We only consider in these experiments our mono-lingual English models, because we have not found any cross- or multi-lingual results in the literature about dialog act recognition to compare with. First, we report the results of traditional feature-engineering methods, which combine a rich set of handcrafted features with dialogue history using Bayesian n-gram [6] or TBL classifier [10]. These methods have obtained the best score on the Verbmobil corpus so far.

We further implemented another baseline that uses a maximum entropy (ME) classifier with simple bag of words (BoW) features. Then, we show the results of our previous LSTM system [18], which uses only simple word level features and word tokens from the previous dialogue act. The results of our approaches are presented in the last two lines of this table.

Although we have done our best to replicate the same experimental set-up as in the related works, some doubts subsist, because the training/testing splits are not available. Therefore, the reported results of the first two methods may not be precisely compared with the others. However, we can still conclude that the performance of our methods is comparable with the state of the art.

6. Conclusions

In this paper, we have proposed and compared several methods for multi-lingual and cross-lingual DA recognition based on deep neural networks. The first approach builds one general model that is trained on the embeddings from all available languages, while the second one trains the model only on one pivot language and cross-lingual projection is achieved by the CCA transform method. We have compared and evaluated two different CNN configurations and one Bi-LSTM on the Verbmobil corpus with English and German DAs.

We have shown that the multi-lingual model significantly outperforms the cross-lingual approach. Another advantage of the multi-lingual model is that it does not need language detection. However, the multi-lingual model is less flexible and may not scale easily to many languages, because retraining is necessary when adding new languages. We have further shown that fine-tuning of word2vec embeddings does not bring any improvement in Verbmobil. We have confirmed that the dialogue history is beneficial for DA recognition in almost all cases and that the best neural classifier is the Bi-LSTM network. We have also compared our approaches with several state-of-the-art methods in mono-lingual scenario and concluded that the performance of our methods is comparable with the state of the art.

The generic approaches depicted in this work may be improved in many ways, e.g., by exploiting contextual word embeddings, transformer-based encoders and by annotating more languages. In the short term, it would also be interesting to improve the multi-lingual model by better handling words that have the same form across languages.

7. Acknowledgements

This work has been partly supported from ERDF ”Research and Development of Intelligent Components of Advanced Technologies for the Pilsen Metropolitan Area (InteCom)” (no.: CZ.02.1.01/0.0/0.0/17_048/0007267) and by Grant No. SGS-2019-018 Processing of heterogeneous data and its specialized applications.

8. References

[1] A. Klementiev, I. Titov, and B. Bhattarai, "Inducing crosslingual distributed representations of words," *Proceedings of COLING 2012*, pp. 1459–1474, 2012.

[2] R. Agerri and G. Rigau, "Robust multilingual named entity recognition with shallow semi-supervised features," *Artificial Intelligence*, vol. 238, pp. 63–82, 2016.

[3] A. Björkelund, L. Hafdell, and P. Nugues, "Multilingual semantic role labeling," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, 2009, pp. 43–48.

[4] S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J. J. Quantz, "Dialogue acts in verbmobil," 1995.

[5] J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel, *Dialogue acts in Verbmobil 2*. DFKI Saarbrücken, 1998.

[6] N. Reithinger and M. Klesen, "Dialogue act classification using language models," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[7] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[8] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–1061.

[9] M. Mast, H. Niemann, E. Nöth, and E. G. Schukat-Talamazzini, "Automatic classification of dialog acts with semantic classification trees and polygrams," in *International Joint Conference on Artificial Intelligence*. Springer, 1995, pp. 217–229.

[10] K. Samuel, S. Carberry, and K. Vijay-Shanker, "Dialogue act tagging with transformation-based learning," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*. Association for Computational Linguistics, 1998, pp. 1150–1156.

[11] M. Rotaru, "Dialog Act Tagging using Memory-Based Learning," University of Pittsburgh, Tech. Rep., Spring 2002, term Project in. Dialog Systems.

[12] P. Král and C. Cerisara, "Automatic dialogue act recognition with syntactic features," *Language resources and evaluation*, vol. 48, no. 3, pp. 419–441, 2014.

[13] E. Shriberg *et al.*, "Can prosody aid the automatic classification of dialog acts in conversational speech?" in *Language and Speech*, vol. 41, 1998, pp. 439–487.

[14] S. Keizer, A. R., and A. Nijholt, "Dialogue act recognition with Bayesian networks for Dutch dialogues," in *3rd ACL/SIGdial Workshop on Discourse and Dialogue*, Philadelphia, USA, July 2002, pp. 88–94.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[16] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[17] J. Y. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2016, pp. 515–520. [Online]. Available: http://aclweb.org/anthology/N16-1062

[18] C. Cerisara, P. Kral, and L. Lenc, "On the effects of using word2vec representations in neural networks for dialogue act recognition," *Computer Speech & Language*, vol. 47, pp. 175–193, 2018.

[19] H. Khanpour, N. Guntakandla, and R. Nielsen, "Dialogue act classification in domain-independent conversations using a deep recurrent neural network," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 2012–2021.

[20] N. Duran and S. Battle, "Probabilistic word association for dialogue act classification with recurrent neural networks," in *International Conference on Engineering Applications of Neural Networks*. Springer, 2018, pp. 229–239.

[21] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard swbd-damsl shallow-discourse-function annotation coders manual (1997)," *URL http://web. stanford. edu/˜ jurafsky/ws97/manual. august1. html.*

[22] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The icsi meeting recorder dialog act (mrda) corpus," in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, 2004.

[23] J.-M. Benedı, E. Lleida, A. Varona, M.-J. Castro, I. Galiano, R. Justo, I. López, and A. Miguel, "Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana," in *Fifth International Conference on Language Resources and Evaluation (LREC)*, 2006, pp. 1636–1639.

[24] P. Král, C. Cerisara, and J. Klečková, "Combination of classifiers for automatic recognition of dialog acts," in *Interspeech'2005*. Lisboa, Portugal: ISCA, September 2005, pp. 825–828.

[25] L. M. R. Barahona, A. Lorenzo, and C. Gardent, "Building and exploiting a corpus of dialog interactions between french speaking virtual and human agents," in *The eighth international conference on Language Resources and Evaluation (LREC)*, 2012, pp. 1428–1435.

[26] S. Mezza, A. Cervone, E. Stepanov, G. Tortoreto, and G. Riccardi, "Iso-standard domain-independent dialogue act tagging for conversational agents," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3539–3551. [Online]. Available: https://www.aclweb.org/anthology/C18-1300

[27] J. Martínek, L. Lenc, and P. Král, "Semantic space transformations for cross-lingual document classification," in *27th International Conference on Artificial Neural Networks (ICANN 2018)*, vol. 11139 LNCS. Rhodes, Greece: Springer International Publishing, October 4-7 2018, pp. 608–616.

[28] T. Brychcín, "Linear transformations for cross-lingual semantic textual similarity," *arXiv preprint arXiv:1807.04172*, 2018.

[29] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.