IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

End-to-End Multi-Level Dialog Act Recognition

Eugénio Ribeiro[1,2]*, Ricardo Ribeiro*[1,3]*, and David Martins de Matos*[1,2]

[1]L[2]F – Spoken Language Systems Laboratory – INESC-ID Lisboa, Portugal
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3]Instituto Universitário de Lisboa (ISCTE-IUL), Portugal

eugenio.ribeiro@l2f.inesc-id.pt

Abstract

The three-level dialog act annotation scheme of the DIHANA corpus poses a multi-level classification problem in which the bottom levels allow multiple or no labels for a single segment. We approach automatic dialog act recognition on the three levels using an end-to-end approach, in order to implicitly capture relations between them. Our deep neural network classifier uses a combination of word- and character-based segment representation approaches, together with a summary of the dialog history and information concerning speaker changes. We show that it is important to specialize the generic segment representation in order to capture the most relevant information for each level. On the other hand, the summary of the dialog history should combine information from the three levels to capture dependencies between them. Furthermore, the labels generated for each level help in the prediction of those of the lower levels. Overall, we achieve results which surpass those of our previous approach using the hierarchical combination of three independent per-level classifiers. Furthermore, the results even surpass the results achieved on the simplified version of the problem approached by previous studies, which neglected the multi-label nature of the bottom levels and only considered the label combinations present in the corpus.

Index Terms: dialog act recognition, DIHANA corpus, multi-level classification, multi-label classification

1. Introduction

Dialog act recognition is an important task in the context of Natural Language Understanding (NLU) since dialog acts reveal the intention behind the uttered words, allowing the application of specialized interpretation approaches. Consequently, it has been widely explored over the years on multiple corpora with different characteristics [1]. In this sense, the distinguishing aspect of the DIHANA corpus [2], which features interactions in Spanish between humans and a train information dialog system, is its three-level annotation scheme. While the top level refers to the generic task-independent dialog act, the others complement it with task-specific information. Additionally, while each segment has a single top-level label, it may have multiple or no labels on the other levels. Thus, the DIHANA corpus poses both multi-level and multi-label classification problems. However, most previous studies on this corpus approached the task as a single-label classification problem in which the label of a segment was the combination of all its labels. Contrarily, in [3] we explored each level independently, approaching the bottom levels as multi-label classification problems, and then combined the best classifiers for each level hierarchically. Among other conclusions, in that study we have shown that there are dependencies between the multiple levels which the independent classifiers cannot capture unless the information is explicitly provided. Thus, in this paper we approach the problem using an end-to-end classifier to predict the labels of the three levels in parallel, so that the relations between the levels are captured implicitly. Additionally, we explore approaches on segment and context information representation which have recently been proved successful on the dialog act recognition task and were not used in our previous study on the DIHANA corpus.

In the remainder of the paper we start by providing an overview of previous work on dialog act recognition on the DIHANA corpus and how it can be improved, in Section 2. Then, in Section 3, we describe our experimental setup, including the corpus, the network variations, and the training and evaluation approaches. The results of our experiments are presented and discussed in Section 4. Finally, Section 5 states the most important conclusions of this study.

2. Related Work

Automatic dialog act recognition is a task that has been widely explored using multiple classical machine learning approaches, from Hidden Markov Models (HMMs) to Support Vector Machines (SVMs) [1]. However, recently, most approaches on the task take advantage of Deep Neural Network (DNN) architectures to capture different aspects of the dialog [4, 5, 6, 7, 8, 9].

Similarly to the studies on English data, the first studies on the DIHANA corpus employed HMMs using both prosodic [10] and textual [11] features. The study using prosodic features focused on the prediction of the generic top level labels, while the study using textual features considered the combination of the multiple levels. Additionally, the latter study, as well as a more recent one [12], explored the recognition of dialog acts on unsegmented turns using n-gram transducers. However, in those cases, the focus was on the segmentation process. The results of the HMM-based approaches were surpassed in a study that applied SVMs [13] to a feature set consisting of word n-grams, the presence of wh-words and punctuation, and context information from up to three preceding segments in the form of the same features. All of these studies neglected the multi-label nature of the bottom levels of the dialog act annotation scheme of the DIHANA corpus and approached a simplified single-label problem in which the label set consisted of the label combinations present in the corpus. However, this approach limits the possible combinations to those existing in the dataset and neglects the distinguishing characteristics of each individual label.

Contrarily to those studies, in [3] we explored each level independently, approaching the bottom levels as multi-label classification problems, and then combined the best classifiers for each level hierarchically. In that study we compared those that were the two top performing DNN-based approaches on segment representation for dialog act recognition on the Switchboard Dialog Act Corpus [14], which is the most explored cor-

10.21437/IberSPEECH.2018-63

pus for the task. One of those approaches uses a stack of Long Short Term Memory (LSTM) units to capture long distance relations between tokens [7], while the other uses multiple parallel Convolutional Neural Networks (CNNs) with different context window sizes to capture different functional patterns [8]. We have shown that the CNN-based approach leads to better results on every level of the DIHANA corpus. However, while wider context windows are better for predicting the generic dialog acts of the top level, the task-specific bottom levels are more accurately predicted when using narrower windows. Furthermore, recently, we have shown that the performance can be improved by using a Recurrent Convolutional Neural Network (RCNN)-based segment representation approach that is able to capture long distance relations and discards the need for selecting specific window sizes for convolution [9]. Additionally, we have shown that a character-based segment representation approach achieves similar or better results than an equivalent word-based approach on the Switchboard corpus and the top level of the DIHANA corpus and that the information captured by both approaches is complementary [15].

In [3] we have also shown that context information concerning the dialog history and the classification of the upper levels is relevant for the task. Concerning the dialog history, we have explored the use of information from up to three preceding segments in the form of their classifications. On the first two levels, similarly to what happened in previous studies on the influence of context on dialog act recognition [16, 8], we have observed that the first preceding segment is the most important and that the influence decays with distance. On the other hand, since the bottom level refers to information that is explicitly referred to in the segment, it is not influenced by information from the preceding segments, at least at the same level. Recently, we have shown that the representation of information from the preceding segments used in previous studies does not take the sequentiality of those segments into account and that the whole dialog history can be summarized in order to capture that information as well as relations with more distant segments [9]. Additionally, in this paper, we further explore the relations between levels by using an end-to-end approach to predict the labels of the three levels in parallel and capture those relations implicitly.

3. Experimental Setup

This section presents our experimental setup, starting with a description of the corpus, followed by an overview of the aspects addressed by our experiments and the used network architecture and a description of the training and evaluation approaches.

3.1. Dataset

The DIHANA corpus [2] consists of 900 dialogs between 225 human speakers and a Wizard of Oz telephonic train information system. There are 6,280 user turns and 9,133 system turns, with a vocabulary size of 823 words. The turns were manually transcribed, segmented, and annotated with dialog acts [17]. The total number of annotated segments is 23,547, with 9,715 corresponding to user segments and 13,832 to system segments.

The dialog act annotations are hierarchically decomposed in three levels [18]. The top level, Level 1, represents the generic intention of the segment, while the others refer to task-specific information. There are 11 Level 1 labels, out of which two are exclusive to user segments and four to system segments. Overall, the most common label is *Question*, covering 27% of the segments, followed by the *Answer* and *Confirmation* labels, covering 18% and 15%, respectively. This is consistent with the information-transfer nature of the dialogs.

Although they share most labels, the two task-specific levels focus on different information. While Level 2 is related to the information that is implicitly focused in the segment, Level 3 is related to the kind of information that is explicitly referred to in the segment. There are 10 labels common to both levels and three additional ones on Level 3. The most common Level 2 labels are *Departure Time*, *Fare*, and *Day*, which are present in 32%, 14%, and 8% of the segments, respectively. On the other hand, the Level 3 label distribution is more balanced, with the most common labels, *Destination*, *Day*, and *Origin*, being present in 16%, 16%, and 13% of the segments, respectively.

While a segment has a single Level 1 label, it may have multiple or no labels in the other levels. In this sense, only 63% of the segments have Level 2 labels, and that percentage is even lower, 52%, when considering Level 3 labels. This is mainly due to the fact that Level 1 labels concerning dialog structuring or communication problems cannot be paired with any labels in the remaining levels.

3.2. End-to-End Neural Network Architecture

In our experiments, we incrementally built the architecture of our network by assessing the performance of different approaches for each step. However, due to space constraints, we are not able to show individual figures for all of those approaches. Thus, in Figure 1 we show the architecture of the final network and use it to refer to the alternatives we explored.

At the top are our two complementary segment representation approaches. On the left is the word-based approach, which captures information concerning both word sequences and functional patterns using the adaptation of the RCNN by Lai et al. [19] which we introduced in [9]. In our adaptation we replaced the simple Recurrent Neural Networks (RNNs) used to capture the context surrounding each token by Gated Recurrent Units (GRUs), in order to capture relations with more distant tokens. To represent each word, in our experiments we used 200-dimensional Word2Vec [20] embeddings trained on the Spanish Billion Word Corpus [21]. On the right is the character-based approach we introduced in [15], which uses three parallel CNNs with different window sizes to capture relevant patterns concerning affixes, lemmas, and inter-word relations. We performed experiments using each approach individually, as well as their combination, which is depicted in Figure 1.

The representation of the segment can then be combined with context information concerning the dialog history and speaker changes. We provide the latter in the form of a flag stating whether the speaker changed in relation to the previous segment, as in [8, 9]. To provide information from the preceding segments we use the approach we introduced in [9], which summarizes the dialog history by passing the sequence of dialog act labels through a GRU. We performed experiments using a single summary combining information concerning the three levels, as well as using per-level summaries which summarize the sequence of preceding labels of each level individually.

To predict the dialog act labels for the segment, the combined representation is passed through two dense layers. While the first reduces its dimensionality and identifies the most relevant information present in that representation, the second generates the labels. In terms of the dimensionality reduction layer, we experimented using a single layer that captures the most relevant information that is generic to the three levels, as well as per-level dimensionality reduction layers, which capture the

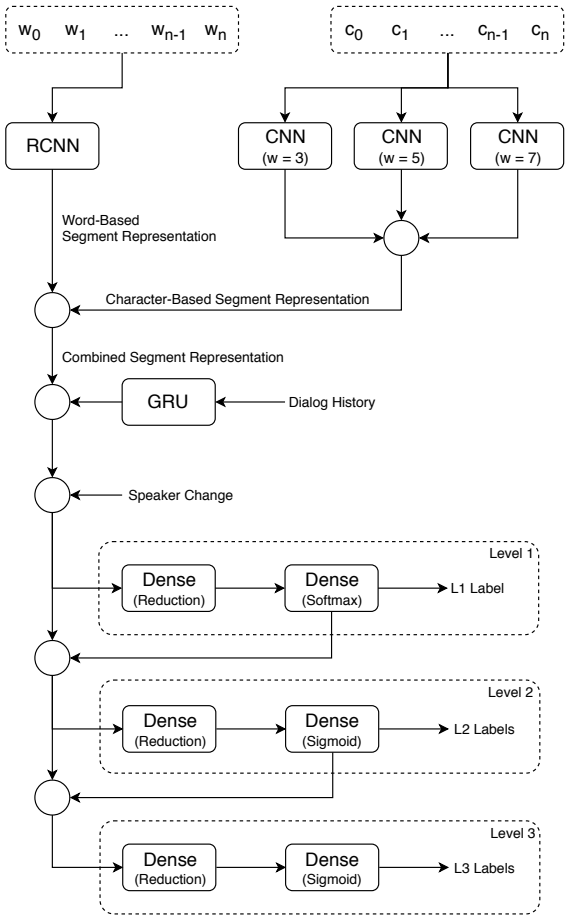

Figure 1: *The architecture of our network. w_i refers to the embedding representation of the i-th word, while c_i refers to the embedding representation of the i-th character. The circles represent the concatenation of the different inputs.*

most relevant information for each level, as depicted in Figure 1. Since the first level poses a single-label classification problem, the output layer uses the softmax activation and the categorical cross entropy loss function. On the other hand, since the other levels pose multi-label classification problems, the corresponding output layers use the sigmoid activation and the binary cross entropy loss function which, given the possibility of multiple labels, is actually the Hamming loss function [22]. In both cases, for performance reasons, we use the Adam optimizer [23].

In [3], we have shown that the prediction of dialog act labels of a certain level is improved when information concerning the upper levels is available. Thus, as shown in Figure 1, we also performed experiments that considered the output from the upper levels in the dimensionality reduction layers.

3.3. Training and Evaluation

To implement our networks we used Keras [24] with the TensorFlow [25] backend. We used mini-batching with batches of size 512 and the training phase stopped after 10 epochs without improvement. The results presented in the next section refer to the average (μ) and standard deviation (σ) of the results obtained over 10 runs. On each run, we performed 5-fold cross-validation using the folds defined in the first experiments on the DIHANA corpus [10, 11]. In terms of evaluation metrics we use accuracy. This metric is penalizing for the multi-label classification scenarios of Level 2 and Level 3, since it does not account for partial correctness [26]. However, due to space constraints and since we are focusing on the combined prediction of the three levels, we do not report results for specialized metrics.

4. Results

In this section we present the results of our experiments on each level, as well as on the combination of the three levels. Both the results on Level 1 and on the combination of the three levels can be directly compared with those reported in [3]. However, that is not the case for the remaining levels. In [3], since we explored each level independently and the annotation scheme does not allow segments with a Level 1 label concerning dialog structuring or communication problems to have labels in the remaining levels, we did not consider those segments when training and evaluating the Level 2 and Level 3 classifiers. Contrarily, since in this study we use a single classifier to predict the labels for all levels, those segments are also considered.

We started by exploring the word- and character-based segment representation approaches, as well as their combination. Thus, in these experiments, we did not provide context information to the network and we used a single dimensionality reduction layer for the three levels. In Table 1 we can see that, as we have previously shown in [15], the character-based approach leads to better results than the word-based one on Level 1. Additionally, the results achieved by the word-based approach are above those reported in [15], which confirms that the word-level RCNN-based segment representation approach leads to better results than the CNN-based one we used in [15] and [3]. On the other hand, the slight performance decrease of the character-based approach can be explained by the combined prediction of the three levels, which does not allow the classifier to specialize in predicting Level 1 labels. On the remaining levels, the character-level approach still performs better. However, the difference is smaller than on Level 1, which is explained by the more prominent relation of the labels of these levels with specific words. Furthermore, the combination of both approaches leads to the best results on every level.

Table 1: *Accuracy (%) results according to the segment representation approach.*

	Level 1		Level 2		Level 3		All	
Approach	μ	σ	μ	σ	μ	σ	μ	σ
Word-Based	92.18	.13	79.36	.18	79.35	.20	75.82	.17
Character-Based	95.31	.07	81.25	.47	81.24	.54	78.67	.51
Combined	95.64	.07	82.46	.20	82.44	.21	79.88	.17

By using per-level dimensionality reduction layers, the classifier is able to select the information that is most relevant for predicting the labels of each level. Thus, as shown in Table 2, this adaptation leads to improved results on the two bottom levels and on the combination of the three levels. However, the performance on Level 1 did not improve, which suggests that the combined segment representation captures more information concerning specific words in detriment of functional patterns relevant for the prediction of some Level 1 labels. Providing information concerning the output generated for the upper levels leads to further improvement, in line with that reported in [3] in spite of not using gold standard labels.

As stated in Section 2, context information from the pre-

Table 2: *Accuracy (%) results according to the dimensionality reduction approach.*

	Level 1		Level 2		Level 3		All	
Approach	μ	σ	μ	σ	μ	σ	μ	σ
Single Reduction	95.64	.07	82.46	.20	82.44	.21	79.88	.17
Per-Level Reduction	95.64	.05	83.21	.11	83.17	.18	80.23	.16
Output Waterfall	95.65	.05	83.29	.21	83.36	.19	80.49	.24

ceding segments has been proved important in many studies on dialog act recognition. The results in Table 3 confirm this importance for all levels. However, there are different conclusions to draw depending on the level. Concerning the first level, we can see that considering the dialog history leads to an average accuracy improvement of 3.72 percentage points, which is above the 3.42 reported in [15]. Considering that the classifier fails to predict the correct Level 1 label for less than one percent of the segments, this is a relevant improvement which is explained by the representation of the dialog history in the form of a summary. In [3] we have shown that the dialog history is not relevant when only the Level 3 is considered, since it refers to information that is explicitly referred to in the segment. However, in Table 3 we can see an average accuracy improvement of 12.86 percentage points on Level 3 when considering the dialog history. This is explained by the fact that the provided information concerns all levels and, as shown in [3], information from the preceding segments concerning the upper levels is relevant when predicting Level 3 labels. On the one hand, what is implicitly targeted at given time is expected to be explicitly referred to in the future. Thus, there is a relation between the Level 2 labels of preceding segments and the Level 3 labels of the current one. On the other hand, the dialogs feature multiple question-answer pairs for which the labels on the lower levels are the same. Thus, when the Level 1 label of the preceding segment is *Question*, the Level 3 labels of that segment are typically present in the current segment as well. This relation between levels is further confirmed by the improved performance when using a single summary for the whole dialog history in comparison to when using independent per-level summaries.

Table 3: *Accuracy (%) results when using context information from the preceding segments.*

	Level 1		Level 2		Level 3		All	
Approach	μ	σ	μ	σ	μ	σ	μ	σ
No Information	95.65	.05	83.29	.21	83.36	.19	80.49	.24
Single Summary	99.37	.01	96.15	.11	96.22	.15	95.53	.14
Per-Level Summary	99.34	.03	95.80	.11	95.87	.13	95.14	.14

As shown in previous studies, using information concerning speaker changes slightly improves the performance, up to 95.64% average accuracy on the combination of the three levels. More importantly, as discussed in [3], the system segments are scripted and, thus, are easier to predict than the user segments. Furthermore, a dialog system is aware of its own dialog acts and must only predict those of its conversational partners. As expected, the performance decreases if the classifier is trained and evaluated on user segments only. The average decrease on the combination of the three levels is of 4.5 percentage points. However, on Level 1 it is of just .67 percentage points.

Since we use a single classifier to predict the labels for the three-levels, there is no explicit restriction that segments with Level 1 labels concerning dialog structuring or communication problems cannot have labels in the remaining levels. However, if we post-process the results to enforce that restriction, the improvement on the combination of the three levels is of just .03 percentage points when considering all segments and .1 when considering user segments only. This shows that the network is able to learn that restriction based on the training examples.

Overall, the average accuracy of our best approach on the combination of the three-levels is 95.67%. This result is 3.33 percentage points above the 92.34% we achieved in [3] using the hierarchical combination of independent classifiers for each level. Furthermore, it is even above the 93.98% achieved when considering the single-label simplification of the problem, which only considers the label combinations present in the corpus. This shows that the network is able to capture relevant relations between levels while still being able to identify the most important information for each level using the per-level dimensionality reduction layers.

5. Conclusions

In this paper we have presented our approach on dialog act recognition on the DIHANA corpus using an end-to-end classifier to predict the labels for the three levels defined in the annotation scheme. This way, the relations between levels are captured implicitly, contrarily to what happened in our previous approach on the task, which used independent per-level classifiers. Additionally, we have used approaches on segment and context information representation which have recently been proved more appropriate for the task.

First, we have shown that character-based segment representation also performs better than word-based representation on the multi-label classification problems and that the combination of both approaches surpasses each individual approach. In this sense, on the combination of all levels, the combined approach surpassed the word- and character-level approaches by around four and one percentage points, respectively.

Then, we have shown that it is important to have per-level dimensionality reduction layers in order to specialize the segment representation for each level. Additionally, the performance is improved when a cue for the hierarchical relation between the levels is provided by considering the output for the upper levels when predicting the labels for each level.

Furthermore, we have shown that the relation between levels is also important when providing context information concerning the dialog history, as a combined summary of the classifications of the preceding segments led to better results than three independent per-level summaries.

Finally, by providing information concerning speaker changes and forcing the segments with Level 1 labels concerning dialog structuring or communication problems to have no labels on the remaining levels, we achieved 95.67% accuracy on the combination of the three levels, which is over three percentage points above our previous approach and even surpasses the results achieved on the simplified single-label classification problem approached by previous studies.

As future work, we intend to explore how our approach can be adapted to perform automatic segmentation of the turns instead of relying on a priori segmentation and assess the impact on the overall dialog act recognition performance.

6. Acknowledgements

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

7. References

[1] P. Král and C. Cerisara, "Dialogue Act Recognition Approaches," *Computing and Informatics*, vol. 29, no. 2, pp. 227–250, 2010.

[2] J.-M. Benedí, E. Lleida, A. Varona, M.-J. Castro, I. Galiano, R. Justo, I. L. de Letona, and A. Miguel, "Design and Acquisition of a Telephone Spontaneous Speech Dialogue Corpus in Spanish: DIHANA," in *LREC*, 2006, pp. 1636–1639.

[3] E. Ribeiro, R. Ribeiro, and D. M. de Matos, "Hierarchical Multi-Label Dialog Act Recognition on Spanish Data," *Traitement Automatique des Langues (submitted to)*, vol. 59, no. 1, 2019.

[4] N. Kalchbrenner and P. Blunsom, "Recurrent Convolutional Neural Networks for Discourse Compositionality," in *Workshop on Continuous Vector Space Models and their Compositionality*, 2013, pp. 119–126.

[5] J. Y. Lee and F. Dernoncourt, "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks," in *NAACL-HLT*, 2016, pp. 515–520.

[6] Y. Ji, G. Haffari, and J. Eisenstein, "A Latent Variable Recurrent Neural Network for Discourse Relation Language Models," in *NAACL-HLT*, 2016, pp. 332–342.

[7] H. Khanpour, N. Guntakandla, and R. Nielsen, "Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network," in *COLING*, 2016, pp. 2012–2021.

[8] Y. Liu, K. Han, Z. Tan, and Y. Lei, "Using Context Information for Dialog Act Classification in DNN Framework," in *EMNLP*, 2017, pp. 2160–2168.

[9] E. Ribeiro, R. Ribeiro, and D. M. de Matos, "Deep Dialog Act Recognition using Multiple Token, Segment, and Context Information Representations," *CoRR*, vol. abs/1807.08587, 2018. [Online]. Available: http://arxiv.org/abs/1807.08587

[10] V. Tamarit and C.-D. Martínez-Hinarejos, "Dialog Act Labeling in the DIHANA Corpus using Prosody Information," in *V Jornadas en Tecnología del Habla*, 2008, pp. 183–186.

[11] C.-D. Martínez-Hinarejos, J.-M. Benedí, and R. Granell, "Statistical Framework for a Spanish Spoken Dialogue Corpus," *Speech Communication*, vol. 50, no. 11–12, pp. 992–1008, 2008.

[12] C. D. Martínez-Hinarejos, J.-M. Benedí, and V. Tamarit, "Unsegmented Dialogue Act Annotation and Decoding with N-Gram Transducers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 198–211, 2015.

[13] B. Gambäck, F. Olsson, and O. Täckström, "Active Learning for Dialogue Act Classification," in *INTERSPEECH*, 2011, pp. 1329–1332.

[14] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual," University of Colorado, Institute of Cognitive Science, Tech. Rep. Draft 13, 1997.

[15] E. Ribeiro, R. Ribeiro, and D. M. de Matos, "A Study on Dialog Act Recognition using Character-Level Tokenization," in *AIMSA*, 2018.

[16] ——, "The Influence of Context on Dialogue Act Recognition," *CoRR*, vol. abs/1506.00839, 2015. [Online]. Available: http://arxiv.org/abs/1506.00839

[17] N. Alcácer, J. M. Benedí, F. Blat, R. Granell, C. D. Martínez, and F. Torres, "Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus," in *SPECOM*, 2005, pp. 583–586.

[18] C.-D. Martínez-Hinarejos, E. Sanchis, F. García-Granada, and P. Aibar, "A Labelling Proposal to Annotate Dialogues," in *LREC*, 2002, pp. 1566–1582.

[19] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification," in *AAAI*, 2015, pp. 2267–2273.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *NIPS*, 2013, pp. 3111–3119.

[21] C. Cardellino, "Spanish Billion Word Corpus and Embeddings," http://crscardellino.me/SBWCE/, 2016.

[22] J. Díez, O. Luaces, J. J. del Coz, and A. Bahamonde, "Optimizing Different Loss Functions in Multilabel Classifications," *Progress in Artificial Intelligence*, vol. 3, no. 2, pp. 107–118, 2015.

[23] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *ICLR*, 2015.

[24] F. Chollet *et al.*, "Keras: The Python Deep Learning Library," https://keras.io/, 2015.

[25] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," https://www.tensorflow.org/, 2015.

[26] M. S. Sorower, "A Literature Survey on Algorithms for Multi-Label Learning," Oregon State University, Tech. Rep., 2010.