Cross-lingual Transfer Learning for Dialogue Act Recognition

Jiří Martinek¹, Christophe Cerisara³, Pavel Král¹,², Ladislav Lenc²

¹ Dept. of Computer Science & Engineering
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
² NTIS - New Technologies for the Information Society
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
³ Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
{ jimar,pkral,llenc}@kiv.zcu.cz, cerisara@loria.fr

Abstract

This paper deals with cross-lingual transfer learning for dialogue act (DA) recognition. Besides generic contextual information gathered from pre-trained BERT embeddings, our objective is to transfer models trained on a standard English DA corpus to two other languages, German and French, and to potentially very different types of dialogue with different dialogue acts than the standard well-known DA corpora. The proposed approach thus studies the applicability of automatic DA recognition to specific tasks that may not benefit from a large enough number of manual annotations. A key component of our architecture is the automatic translation module, which limitations are addressed by stacking both foreign and translated words sequences into the same model. We further compare both CNN and multi-head self-attention to compute the speaker turn embeddings and show that in low-resource situations, the best results are obtained by combining all sources of transferred information.

Index Terms: dialogue acts recognition, cross-lingual transfer learning

1. Introduction

Automatic dialogue act (DA) recognition has reached near-human performance on standard corpora, such as the Switchboard Dialogue Act corpus. However, various application domains lead to different types of dialogues: this variability, which is illustrated in the French corpus described in Section 5.2, severely impacts the direct application of a model trained on a standard corpus to many other types of application tasks, as show next. Furthermore, developing DA-recognition models in other languages than English requires the costly annotation of large-enough corpora. In the last years, efforts have been dedicated to address this issue by proposing language-independent ISO standard [1], but variability in the types of dialogues makes the requirement to annotate dedicated training corpora still pressing. We propose to investigate cross-lingual transfer learning methods to reduce the amount of annotations required for each new domain and language.

Transfer learning aims at reusing knowledge gained from a large corpus to improve the performances of small models trained on a related task with low resources. We investigate in this work two sources of information from which we transfer knowledge: pre-trained English BERT sentence (vectors) embedding, and an English corpus annotated with dialogue acts. Two low-resource target tasks are considered: dialogue act recognition in German and in French, where the amount of annotated dialogue acts is limited to a few hundred samples that may be annotated by one application developer within a few hours. Two target conditions are also considered: either with the same set of dialogue acts in German and in English, or with a different set of dialogue acts in French.

In addition to the relatively large resources available in English, we further assume the availability of a relatively good machine translation system; we will use for this purpose Google’s translation. Hence, our transfer learning strategy consists first in translating every target training and test material to English, and then fine-tuning our initial “large” English model to recognise these translated dialogues. Details of this method are given in Section 3.

2. Related Work

New approaches in the dialogue act recognition field are mainly evaluated on English datasets. Some methods have also been tested on other languages, such as Spanish (DIHANA corpus [2]), Czech [3], French [4] and German (Verbmobil [5] corpus). A nice comparison and state-of-the-art results on different DA datasets is summarised in [6]: 82% accuracy is reached on the Switchboard Dataset (SwDA) [7], 80% on the German part of the Verbmobil (VM) corpus and almost 90% on the MRDA (Meeting Recorder Dialog Act) dataset [8]. Nevertheless, different DA labels in datasets is an obstacle for effective multi-lingual and multi-dataset research. Several interesting research efforts have thus emerged to define and exploit generic dialogue acts [1]. We rather focus next on specific types of dialogues and task-related dialogue acts.

Transfer learning has been quite a popular approach in deep learning in recent years. Such approaches help to speed up the training process and are popular particularly in computer vision tasks (e.g. automated pavement distress detection and classification [9] or a transfer learning algorithm for face verification [10]). A well-known easy transfer learning approach in the natural language processing (NLP) field is the usage of pre-trained word2vec (W2V) [11] or other word embeddings.

In standard transfer learning information flows from the
source to the target domain (one direction only). A related approach is multi-task learning \[12\] where information flow across all tasks (usually more than two) because what is learned for each task can help other tasks to be learned better.

Pre-trained contextual embeddings such as ELMo (Peters et al. \[13\]) and in particular BERT (Devlin et al. \[14\]) are essential in many NLP tasks. In the dialogue act recognition domain, Dai et al. \[15\] fine-tuned BERT model to classify a single utterance with quite good results. Wu et al. \[16\] propose task-oriented dialogue BERT (ToD BERT). A survey of deep DA recognition using multiple representations is made by Ribeiro et al. \[17\]. BERT embeddings led to better results than other representations.

3. Models

3.1. English DA Classifier

Our initial English dialogue act recognition model trained on the English dataset annotated with dialogue acts is a multi-layer perceptron (MLP) with BERT embeddings as inputs. We have tested various topologies for this initial model, and chosen this one because of its good performances and fast training times.

Each speaker turn (utterance), composed of a variable number of words, is first encoded into a single pre-trained 1024-dimensional sentence embedding vector with BERT Large\[1\].

As shown in Figure 1, two such vectors are actually computed, respectively for the previous and the current speaker turns, and concatenated as an input to the MLP. The MLP outputs 17 tags, which correspond to the dialogue act labels of the Verbmobil-EN corpus. This MLP has been trained on the training part of the Verbmobil-EN corpus.

3.2. Speaker Turn Embeddings

In our MLP model described previously, variable length sentences are encoded into a unique speaker turn embedding vector with a pretrained BERT model. We have further used two other models to compute the speaker turn embeddings: a convolutional neural network (CNN) and a multi-head self-attention (MH-SAtt) model.

The CNN model, shown in Figure 2 is derived from the model of Martinek et al. \[18\]. It takes as an input a word sequence truncated/padded to 15 words that always include the last two words of the sequence, following \[19\]. These 15 words are encoded into either random embeddings or W2V vectors. The CNN outputs a 256-dimensional vector for the current speaker turn.

The MH-SAtt model transforms each input random word embedding with the standard scaled dot-product multi-head self-attention module \[20\]. A global max pooling operation is then applied to compute the speaker turn embedding. Figure 3 shows how this model is used in our experiments.

In all our models, the previous speaker turn is also encoded into a 1024-dimensional vector with BERT and injected into the classification step: this is an easy way to take into account the previous dialogue act, without requiring a costly RNN, CRF and/or beam search procedure.

4. Transfer Learning Approach

Pre-trained BERT are famous for transferring general English lexical information into a large variety of downstream NLP tasks. We propose an approach to exploit them for cross-lingual dialogue act recognition through automatic language transla-

\[1\]from https://github.com/google-research/bert#pre-trained-models

\[1\] https://github.com/CyberZHG/keras-multi-head
5.1. German Data
We build a low-resource German dialogue act corpus by randomly sampling 100 utterances from the VerbMobil-GE training corpus with the same dialogue acts distribution as in the complete training corpus, see Table 1. The model trained on this small corpus will be tested on the complete (1460 utterances), standard VerbMobil-GE test corpus.

Table 1: Distribution of DAs in VerbMobil-GE corpus

| Label                  | Occurrence | Label                  | Occurrence |
|------------------------|------------|------------------------|------------|
| FEEDBACK               | 28%        | DELIBERATE             | 3%         |
| SUGGEST                | 19%        | INTRODUCE              | 2%         |
| INFORM                 | 18%        | COMMIT                 | 1%         |
| REQUEST                | 9%         | CLOSE                  | 1%         |
| GREET                   | 4%         | POLIT FORM             | 1%         |
| BYE                    | 4%         | THANK                  | 1%         |
| INIT                    | 4%         | DEFER                  | 1%         |
| BACKCHANNEL             | 3%         | OFFER                  | 1%         |

5.2. French Data
We manually annotate 470 turns from the TCOF corpus with dialogue acts. This corpus contains manual transcriptions of spoken dialogues in French recorded by linguists in real-life situations involving volunteer citizens. The types of dialogue in our 470 turns are very different from the ones found in standard DA corpora: they involve two friends trying to find a suitable gift, three students talking about their courses while chatting with someone else on their smartphone and an adult talking with a young girl who is drawing. Hence, the French corpus is annotated with a slightly different set of labels, which distribution is shown in Table 2. This French corpus is freely distributed with a CC BY-NC-SA license.

Table 2: Distribution of DAs in the French corpus

| Label                       | Occurrence | Label                       | Occurrence |
|-----------------------------|------------|-----------------------------|------------|
| INFORM                      | 20%        | OPEN ANSWER                | 7%         |
| AGREE                       | 13%        | DISAGREE                   | 5%         |
| BACKCHANNEL                 | 10%        | YES ANSWER                 | 5%         |
| Y/N QUESTION                | 10%        | NO ANSWER                  | 5%         |
| OPEN QUESTION               | 9%         | OTHER ANSWERS              | 1%         |
| PERFORMATIVE                | 8%         | GREETINGS                  | < 1%       |

5.3. Initial Phase: English Model
The initial English MLP model is trained on the full VerbMobil-EN corpus. The hyper-parameters of this model are tuned on the English corpus: this model is thus trained for 200 epochs with a learning rate of 0.002. Table 3 compares its accuracy when trained either from BERT embeddings or from speaker turn embeddings obtained with the CNN model. The CNN may use initial random or pretrained W2V word embeddings. Every experiment is run 10 times and the results are averaged.

5.4. Fine-Tuning Phase
Fine-tuning consists in training the last classification layers of the models presented in Section 5 on the small foreign corpus that has been translated into English. This fine-tuning process thus involves some additional hyper-parameters, such as the learning rate and the fixed number of epochs, which are tuned on a German development corpus composed of 10,000 utterances randomly sampled from the VerbMobil-GE training

5. Experiments
The objective of the following experiments is to validate our cross-lingual transfer learning proposal for dialogue act recognition on low-resource application domains. We apply transfer learning from English to German and French languages. Every experiment is run 10 times and the results are averaged. The standard deviation is also computed.
Table 3: Initial Phase – English MLP Model trained on Verbmobil-EN (9599 turns) and tested on Verbmobil-EN (1460 turns). The Embeddings column indicates how the word embeddings are initialised.

| Model | Embeddings | Epochs | Test Acc | Std. Dev. |
|-------|------------|--------|----------|-----------|
| MLP   | BERT       | 200    | 0.734    | 0.010     |
| CNN   | W2V        | 200    | 0.704    | 0.009     |
| CNN   | Random     | 200    | 0.702    | 0.008     |

corpus. The same hyperparameter values found on this German corpus are also used for the French model.

5.5. Baseline Approaches

Three baseline classifiers are shown in the first part of Table 4:

1. **Majority Class Classifier (MC)**: it always predicts the most common DA in the training dataset.

2. **Training from Scratch**: Instead of transferring the model parameters from the initial English MLP model, the parameters (without the embeddings) are randomly initialised in this baseline.

3. **No-Fine-Tuning**: The model parameters are simply transferred from the initial English MLP model, and no further training is done.

5.6. Fine-Tuning Experiments

5.6.1. German Results

Table 4 shows the results of our models on the Verbmobil-GE test set. All these models process only translated English sentences, except for the last MH-SAtt model that further includes the original German words, as shown in Figure 3.

Table 4: Accuracy on the Test Verbmobil-GE corpus

| Model | Embeddings | Epochs | Acc | Std. Dev. |
|-------|------------|--------|-----|-----------|
| Baseline MC | – | 0.270 | – | – |
| From scratch | CNN | W2V | 15 | 0.410 | 0.012 |
| | MLP | BERT | 25 | 0.468 | 0.008 |
| Non-Fine-tuning | CNN | W2V | – | 0.380 | – |
| | MLP | BERT | – | 0.479 | – |
| Fine-Tuning | CNN | W2V | 15 | 0.463 | 0.008 |
| | MLP | BERT | 25 | 0.484 | 0.025 |
| | MH-SAtt | BERT + GE | 50 | 0.502 | 0.012 |

The relative contributions of the various sources of information and models are similar to the German experiments. However, the differences between the results are much smaller, which is likely due to the fact that the hyperparameters have not been tuned on French but on German.

6. Conclusions

We have investigated two types of transfer learning: pretrained word embeddings and classifier fine-tuning. Three types of sentence representations, namely BERT, CNN and multi-head self-attention were utilized. The objective was to port a reference DA recognition model trained on English to another language and dialogue context with only a limited amount of annotated resources. We have validated these approaches on two target languages, German and French, and developed a dedicated French DA corpus with real-life dialogues recorded in quite different conditions than the existing standard DA corpora. The main conclusion is that all available sources of information are required to obtain the best results, but also that some data should be reserved for proper hyperparameter tuning. Another conclusion is that automatic translation introduces some bias and mistakes that need to be compensated with fine-tuning. The reduced success of transfer learning for French is likely due to the greater mismatch between the corresponding corpus types, and transfer learning does not totally compensate for the lack of training data in the target domain. Nevertheless, the proposed transfer learning architecture gives good enough results to open the way to future research in transferring rich English DA systems to other languages and less explored domains.

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_048/00072677) and by Grant No. SGS-2019-018 Processing of heterogeneous data and its specialized applications and by the Lorraine Universit´e d’Excellence OLKi project.

8. References

[1] 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](https://www.aclweb.org/anthology/C18-1300)

[2] J.-M. Benedì, 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: Drahana,” in *Fifth International Conference on Language Resources and Evaluation (LREC)*, 2006, pp. 1636–1639.

[3] P. Král, C. Cerisara, and J. Klečková, “Combination of classifiers for automatic recognition of dialog acts,” in *Interspeech* 2005, Lisbon, Portugal: ISCA, September 2005, pp. 825–828.

[4] 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.

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

[6] E. Ribeiro, R. Ribeiro, and D. M. de Matos, “A multilingual and multidomain study on dialog act recognition using character-level tokenization,” *Information*, vol. 10, no. 3, p. 94, 2019.

[7] D. Jurafsky, “Switchboard swbd-damsl shallow-discourse-function annotation coders manual,” *Institute of Cognitive Science Technical Report*, 1997.

[8] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke et al., “The icsi meeting corpus,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03).*., vol. 1. IEEE, 2003, pp. I–I.

[9] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, and A. Agrawal, “Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection,” *Construction and Building Materials*, vol. 157, pp. 322–330, 2017.

[10] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun, “A practical transfer learning algorithm for face verification,” in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3208–3215.

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

[12] R. Caruana, “Multitask learning,” *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” *arXiv preprint arXiv:1802.05365*, 2018.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[15] Z. Dai, J. Fu, Q. Zhu, H. Cui, Y. Qi et al., “Local contextual attention with hierarchical structure for dialogue act recognition,” *arXiv preprint arXiv:2003.06044*, 2020.

[16] C.-S. Wu, S. Hoi, R. Socher, and C. Xiong, “Tod-bert: Pre-trained natural language understanding for task-oriented dialogues,” *arXiv preprint arXiv:2004.06871*, 2020.

[17] E. Ribeiro, R. Ribeiro, and D. M. de Matos, “Deep dialog act recognition using multiple token, segment, and context information representations,” *Journal of Artificial Intelligence Research*, vol. 66, pp. 861–899, 2019.

[18] J. Martínek, P. Král, L. Lenc, and C. Cerisara, “Multi-Lingual Dialogue Act Recognition with Deep Learning Methods,” in *Interspeech 2019*, Graz, Austria, 15–19 September 2019, pp. 1463–1467.

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

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[21] ATILF, “Tcof : Traitement de corpus oraux en français,” 2018, ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr. [Online]. Available: [https://hdl.handle.net/11403/tcof06/2](https://hdl.handle.net/11403/tcof06/2)