Transfer learning in hierarchical dialogue topic classification with neural networks [*]

1st Cesar Montenegro
Intelligent System Group
University of the Basque Country UPV/EHU
Donostia/San Sebastian, Spain
cesar.montenegro@ehu.eus

2nd Roberto Santana
Intelligent System Group
University of the Basque Country UPV/EHU
Donostia/San Sebastian, Spain
roberto.santana@ehu.eus

3rd Jose A. Lozano
Intelligent System Group
University of the Basque Country UPV/EHU
Donostia/San Sebastian, Spain
Basque Center of applied mathematics(BCAM)
Bilbao, Spain
ja.lozano@ehu.eus

April 30, 2021

Abstract

Knowledge transfer between tasks can significantly improve the efficiency of machine learning algorithms. In supervised natural language understanding problems, this sort of improvement is critical since the availability of labelled data is usually scarce. In this paper we address the question of transfer learning between related topic classification tasks. A characteristic of our problem is that the tasks have a hierarchical relationship. Therefore, we introduce and validate how to implement the transfer exploiting this hierarchical structure. Our results for a real-world topic classification task show that the transfer can produce improvements in the behavior of the classifiers for some particular problems.

[*] Funded by the European Commission H2020-SC1-2017-RIA, grant number 769872.

topic classification, transfer learning, hierarchical classification, neural networks

1 Introduction

While for many domains it is usually possible to obtain a set of labelled data that allows the implementation of supervised classification methods, there are situations where this data is scarce or costly to obtain. One of these problems is dialogue topic classification for cases where the topics are very specific, such as dialogues centered on the well-being of seniors [24]. In this natural language understanding (NLU) application domain, annotating the dialogues is a time-consuming and costly process. Unfortunately, the power of the deep neural networks (DNNs) usually applied to address these problems, critically depend on the amount of data.

In the NLU domain, multiple taxonomies have been developed to annotate datasets from different text sources, such as human to human written conversations, telephone transcripts or human to machine interaction [17][1][2][16]. These taxonomies share the strategy of defining a label set with hierarchical relationships, which is used to categorize the information that each particular application needs to recognize. Therefore, hierarchical text classification aims at classifying text sentences or documents into classes that are organized into a hierarchy [11]. Such hierarchy can be structured using a tree, which represents the interrelationships among the classes that share ancestral nodes [12]. The downside of this class structure is that the closer we get to the terminal nodes of the trees, the fewer the instances we have left to learn a model able to distinguish between classes.

One solution to mitigate the limitation of labelled data, is using the parameters learned by a model on an external labeled dataset as starting point to train a model for our classification task. This solution requires that the external labeled dataset must have a label set similar to ours, but depending on the label set of our classification problem, this can be impossible. Using weakly-supervised strategies has been also evaluated to try to mitigate this issue, generating pseudo documents from weakly supervised sources for better model generalization [11]. In this paper we investigate another strategy based on the transfer learning approach [10, 15]. Transfer learning algorithms have shown excellent results in a variety of fields for Machine Learning (ML) applications such as reinforcement learning [23], brain signal analysis and decoding [14, 19], Natural language processing [18], etc. Transfer learning has also been used in NLU tasks such as Named-Entity Recognition (NER) [8] with successful results, although in that particular problem the labels are not structured as a hierarchy.

We focus on a hierarchical dialogue topic classification task in which an utterance can be classified in different classes that are organized in a hierarchical way. The rationale of our approach is to evaluate the transferability of models learned to classify the different tasks involved. We also evaluate the gains that retrained hierarchically related models can produce in the overall performance of the model.

The paper is organized as follows: In the next section we introduce the NLU problem addressed in the paper, and explain its hierarchical structure. In Section 3 we describe the neural network models that are used to address each individual topic classification problem. Section 4 describes the transfer learning strategy implemented. Section 5 presents the experimental framework and Section 6 discusses the results of the experiments. We conclude the paper in Section 7, and also present a number of lines for future research.

2 Hierarchical dialogue topic classification

The work presented in this paper is framed within a multi-disciplinary project that involves the solution of multiple ML tasks. We therefore present a brief introduction to the project to contextualize how data has been collected, labelled and the motivation for the hierarchical label structure.

The main objective of the EMPATHIC Research & Innovation project [24] is to improve the life quality of independent elderly people. In order to achieve this goal, a Personalized Virtual Coach (VC) will engage the users to take care of their diet, to have adequate physical activity, to maintain an active social life and take care of potential chronic diseases. The research problems that need to be solved include not only ML problems related to NLP, but also to the interpretation of body expression and the psychological impact of the physical appearance and gestures of the VC.

An important component in the architecture of a spoken dialogue system is the Dialogue Manager (DM), which maintains the state and manages the flow of the conversation with the user. The decision making of what action must be performed at each turn in order to achieve the coaching objectives is based on the information that the VC is able to obtain from audio and video information from the user, combined with external sources of information about the weather and social or leisure events. A key source of information is the interpretation of the users speech, and for that reason a dialogue act taxonomy is proposed in Montenegro et al. [12] . This taxonomy is composed of three types of labels, namely intent, topic and entities. In this paper we focus on topic classification, this is, we will use the *Topic* label, which assigns, to each utterance, a relevant label that determines the general context in which the conversation is framed. The DM needs to track the topic of the conversation to detect any possible shift and adapt to it. In the work presented by Montenegro et al. [12], a hierarchical structure for the topic labels is proposed. The tree structure for the labels means that an utterance is labelled by tags that can be ordered from more general to more specific. In this structure, the closer a label is to a terminal node, the more precise it is, while the further away it is from the terminal nodes, the more general. Four main groups descend from the root node: *nutrition*, *sport and leisure*, *family engagement* and *other*. Each of these groups further splits into more detailed categories.

Hierarchical classification can be faced with different strategies [21] such as:

- **Flat classification**: a model is trained to predict only classes in the terminal nodes, ignoring the hierarchical structure.
- **Local classifier per level**: creates a model for each level of the hierarchy. It is the least used in the literature.
- **Local classifier per node**: consists of training one binary classifier for each node of the class hierarchy (excluding the root node) solving them as 1-vs-all problems.
- **Local classifier per parent node**: creates a multiclass classifier for each node that has child nodes, to classify between them.
- **Global hierarchy**: a model that learns the whole class hierarchy, and makes a prediction for all nodes at once.

In this paper we follow the "Local classifier per parent node" approach.

3 Neural models for topic classification

Recent trends in NLU have shown a gradual shift to Deep learning models [6] due to their performance when trained with large datasets. More precisely, recurrent neural networks such as long short-term memory (LSTM) [20][22], Gated Recurrent Units (GRU)[3] and Convolutional neural networks (CNN) have been proven to be a very effective approach for a variety of NLU related problems [5][26][9][4]. Topic classification can be faced using multiple types of classifiers and architectures. In this paper we investigate models based on a LSTM based network, and instead of training an ad hoc word-embedding layer, we will use the pretrained embeddings available for Spanish and English from the Spacy library[1]. The decision of using pretrained embeddings is due to the reduced size of the Empathic dataset, and the lack of other resources from similar topics and type of interaction. Moreover, using this external resource allows us to evaluate the influence of the transfer learning approach we propose, isolated from other factors that influence the results. Therefore, the models that will be used for each of the classification tasks of the experimentation, will consist of an LSTM layer followed by a Dense layer with *Relu* activation functions, and finally the output layer with *Sigmoid* activation functions. This architecture is illustrated in Figure 2.

4 Hierarchical transfer learning for dialogue topic classification

We follow the definition of transfer learning given in the survey from Pan and Yang [15]: *"Transfer learning aims at performing a task on a target dataset*

[1] https://spacy.io/

using some knowledge learned from a source dataset". Transfer learning has been proven to be very useful in different deep learning tasks. In addition to mitigating the lack of training data, it also reduces training time and improves performance. Nevertheless, it is difficult to find datasets that have similar labels to those created for a particular task such as the Empathic project, and generating them is a costly and time consuming process. Therefore, we decided to explore other approaches. In the literature of text classification there are two main groups of parameters that are usually transferred. The first group is the one related to the word vectorization mechanism, usually an embedding layer that can be trained with cross domain texts, even if they are not labelled [25]. The second group is more particularized for each problem, as it is the group of parameters that learn what sequences are relevant for the classification task. In order to make transfer learning for this second group of parameters, similarly labelled datasets are required. This is often hard to find, consequently some works try to generate their own new labelled dataset by weakly labeling external datasets [11]. In this work we propose a transfer learning mechanism for the second group of parameters, suitable for classifications tasks where labels have a hierarchical structure, and are compatible with other transfer learning methods.

Tables 1 and 3 illustrate a disadvantage of hierarchical problems, the deeper a classification task is in a hierarchy, the fewer the instances to train a model for that specific task. As an example, Table 1 describes the WOS dataset [7], and the *topic* task, which is the root node, contains 11,967 instances, but the *topic_0* classification task, only one level below the root level, has only 1,498 instances. Also, the lower in the hierarchy a classification problem is, the more specific it becomes, making the task of finding suitable external datasets even more difficult.

The method introduced in this paper exploits an advantage of hierarchical problems over other classification scenarios, and consists of transferring learned parameters between the models of the different classification tasks within the hierarchy. In some cases, the transfer is made from a model situated above in the hierarchy, adding information from a more general classification task, in other cases the information will be transferred from a more specific task, or even from a task which does not add useful information a priori from the topic point of view, but it may transfer key morphological information.

4.1 Notation

In order to describe the experiments accurately, we introduce some notation. Having a hierarchical classification problem like the one illustrated in Figure 1. We denote by $M(t, \theta)$ the model trained to solve task t, initialized with the θ parameters, and $M(t, \varnothing)$ as the model trained to solve task t with random initialization of the parameters. Let $T = t_1, ...t_k$ be the classification tasks, and let $\Theta = \theta_1, ...\theta_k$ be the set of transferable parameters, where θ_i represent the set of transferable parameters obtained from the model $M(t_i, \varnothing)$.

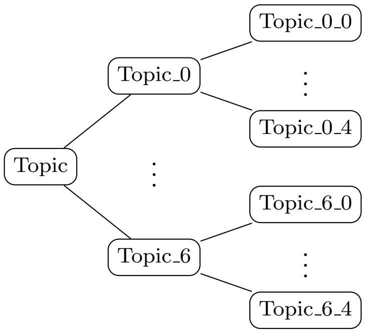

Figure 1: Representation of the hierarchical label structure of the WOS dataset.

5 Experiments

This section describes the experimental setup designed to evaluate the influence of hierarchical transfer learning. For this purpose, we will perform a set of experiments on two hierarchical classification problems using a typical hierarchical dataset (WOS dataset) and the dataset of the Empathic project. Each dataset is labelled with a hierarchical set of labels, and following the Flat classification strategy described in Section 2 we will generate multiple classification tasks for each of them. The models that will be trained to solve each of the classification tasks, will have the LSTM neural network architecture illustrated in Figure 2. The transfer learning method will consist on transfering the parameters belonging to the LSTM layer of the architecture described.

Being the list of classification tasks $T = t_1, ...t_k$, we will train the baseline the models

$$baseline\_models = M(t_1, \varnothing), ...M(t_k, \varnothing)$$

from where we will extract the transferable parameters $\Theta = \theta_1, ...\theta_k$. Then, we will train a model for each task in T initialized with each of the possible transferable parameters in Θ.

$$models\_with\_transfer = \forall t \in T, \forall \theta \in \Theta : M(t, \theta)$$

We will perform a 5-fold cross validation strategy to evaluate the performance of the models for classification task. The performance will be analyzed measuring the F1 score instead of the accuracy to avoid the deceptive results of imbalanced classification tasks. The hypothesis of this paper is that this transfer of parameters should be always beneficial when a model receives parameters from a more general node of its branch, that is, an ancestral node. To test this statement we have performed the experiments described in this section on two different datasets.

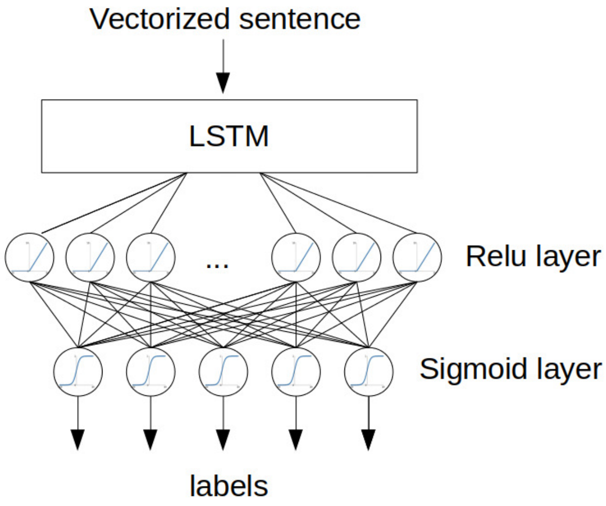

Figure 2: Network structure for topic classification.

5.1 Datasets

The nature of the two datasets chosen for this paper is completely different, despite being two hierarchical topic classification datasets. Their differences make the 5-fold generation procedure to be particularized for each dataset. Next, each dataset and its 5-fold generation procedure is described.

5.1.1 Web of Science dataset

The Web of Science (WOS) [7] dataset is a dataset[2] available online, composed of abstracts extracted from scientific papers of several topics. For this experiment, we have used the WOS-11967 version, containing 11,967 documents with 35 categories, which include 7 parents categories. This dataset is well balanced, and a normal 5-fold strategy will be performed. A description of the number of instances for each class and the tasks to be solved for this dataset can be found in Table 1.

5.1.2 Empathic dataset

The Empathic dataset is the result of the labelling process with the taxonomy described in Montenegro et al. [12] of the transcription of the coaching sessions performed on 72 senior volunteers. These sessions were conducted using a Wizard of Oz procedure to simulate human-machine interactions. The Empathic dataset is multilingual [13], but for the experiments considered in this paper we

[2] https://data.mendeley.com/datasets/9rw3vkcfy4/2

Table 1: Topic classification sub-tasks of the WOS dataset

Node	**#outputs**	**#instances per output label**
Topic	7	1498, 1132, 1959, 1925, 2107, 1617, 1728
Topic_0	5	297, 301, 300, 300, 300
Topic_1	5	300, 0, 353, 53, 426
Topic_2	5	389, 397, 391, 394, 388
Topic_3	5	371, 402, 346, 420, 386
Topic_4	5	410, 423, 384, 441, 449
Topic_5	5	309, 357, 368, 321, 262
Topic_6	5	351, 340, 401, 335, 301

Table 2: Description of the annotated corpus.

Characteristics	**Number**
Number of users	72
Number of dialogues	142
Number of turns	4522
Number of running words	$72,350$
Vocabulary size	5543
Number of topic labels	55

used only Spanish dialogues. Some metrics describing this dataset can be found in Table 2.

Each of these 72 subjects has a particular way of speaking, in order to be rigorous and avoid overfitting to these speech particularities, the sentences belonging to the sessions of each user will not be shared between the training and test sets of any fold. Therefore, to generate the 5 fold cross-validation, the splits will be performed at user level instead of sentence level.

As described in Section 2, a model for each internal node will be trained. In Table 3, the list of internal nodes, number of labels per task, and the amount of instances for each label is detailed. This dataset is highly unbalanced, and some of the tasks have very few instances to benefit from the deep learning power, this is why transfer learning can be useful.

The user split mechanism to create the 5 folds makes this problem even a bigger issue, since not all users talked about the same topics, creating circumstances where the training or test set might not have sentences from some topics. For this reason, instead of the 17 tasks that we should analyze with the ”Local classifier per parent node” strategy, only 12 can generate the training and test sets needed to evaluate the models.

Table 3: Topic classification sub tasks of the Empathic dataset

Node	#labels	#instances per label
topic	4	69, 1048, 8313, 1233
topic_sportandleisure	8	370, 66, 35, 21, 72, 101, 74, 374
topic_sportandleisure_demotivation	3	2, 2, 10
topic_sportandleisure_motivation	3	16, 31, 11
topic_sportandleisure_hobbies	4	8, 54, 16, 183
topic_sportandleisure_sport	4	16, 2, 24, 1
topic_sportandleisure_physicalform	5	17, 6, 3, 10, 8
topic_nutrition	3	185, 193, 112
topic_nutrition_quantity	3	36, 60, 46
topic_nutrition_regularity	2	22, 136
topic_nutrition_variety	3	37, 25, 10
topic_familyandcaregivers	3	15, 3, 8

6 Results

In order to evaluate the transfer learning mechanism proposed, we will examine the results obtained for the WOS and Empathic datasets in terms of F1 score. The results are represented with 3 types of plots for each dataset.

In Figure 3, the results for the WOS dataset are illustrated as a matrix where the vertical axis represents the task for which the model has been trained, and the horizontal axis represents the model from where the transfer has been made. The main purpose of this plot is to perform a visual preliminary analysis of the influence of the transfers.

It is possible to appreciate in Figure 3 that the *topic* task does not get much benefit from any transfer, this is expected since the parameters that are being incorporated to the model were trained with the same data. In other words, the transfer does not incorporate new information to the training procedure in that case. This fact is supported by Figure 4, where the F1-score results for the *topic* tasks are plotted. In this bar plot the black segment represents the baseline result, that is to say, without transfer learning. The performance with transfer learning from the root node, in this case *Topic*, is represented with a gray segment, and the rest of the possible transfers are represented with green segments in the case of the WOS dataset. Therefore, the improvement achieved by transfer learning is not significant in the *topic* task, nevertheless, we can see how the transfer of the parameters learned by the model trained for the root node helps every other task to improve over the baseline results. This result is expected, since, due to the transfer, the models start training in an advantaged position, with information from a more general problem that includes its own problem. The training enables the transferred parameters to specialize in one part of the information hosting network.

Although the *topic* transfer is useful to improve every other task, it is not the

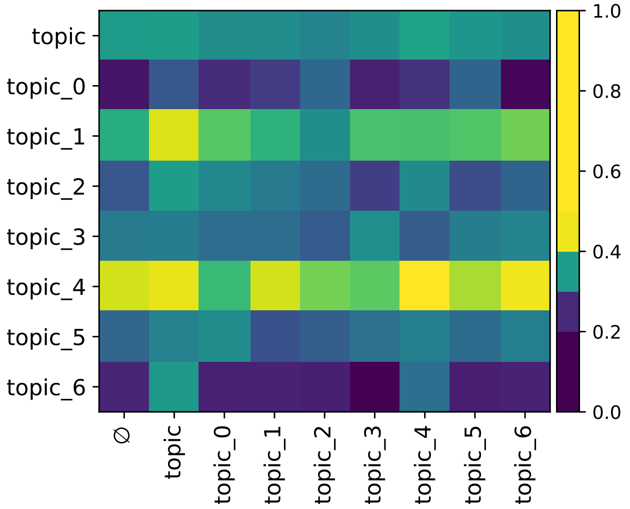

Figure 3: F1 score matrix for every combination of task and transfer in the WOS dataset.

best possible transfer in 4 out of 7 tasks. There are others transfers that have achieved better results, as can be seen in detail in Figure 5, or in Figure 3 in a more general perspective. Analyzing other tasks and transfers, we do not find a clear pattern. In some cases the best transfers come from one particular model, and in other cases from another, but there is always a transfer that improves the performance of the classifier.

Figure 5 shows how, for some tasks, almost any transfer makes the performance improve. We can see how there is only one transfer achieving worse results than the baseline, while the other 7 improve the performance in different rates. On the other hand, in Figure 6 we see how the task only benefits from particular transfers, this fact could be explained if *topic_6* texts are less related to the texts from other topics in terms of topic or writing style. These results suggest that every transfer combination is worth evaluating.

Results for the Empathic show a different behaviour as can be seen in Figure 7. There is not an individual transfer that makes every other task improve. Nevertheless, the root task has a similar response to transfers as in the WOS dataset, that is, there is barely any improvement from any transfer. This fact could be explained because this is the task for which more data is available and models learned with more specific datasets do not seem to produce gains.

Although a transfer that makes every other task improve does not exist, the parameters transfered by the topic_sportandleisure task are the ones that make more tasks improve. This transfer makes 6 other tasks increase their performance, does not significantly affect other 2 tasks, while it makes 3 worse. A probable factor why this task is the most beneficial, is the amount of sen-

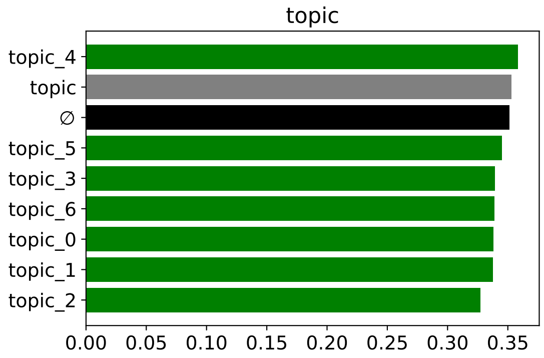

Figure 4: F1 score results for the WOS main topic task.

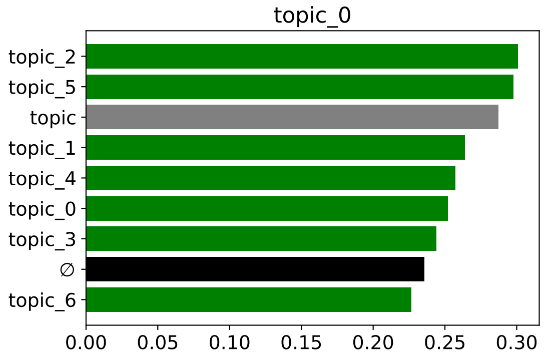

Figure 5: F1 score results for the WOS topic_0 task.

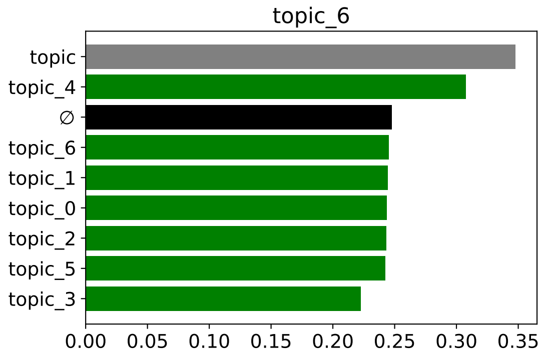

Figure 6: F1 score results for the WOS topic_6 task.

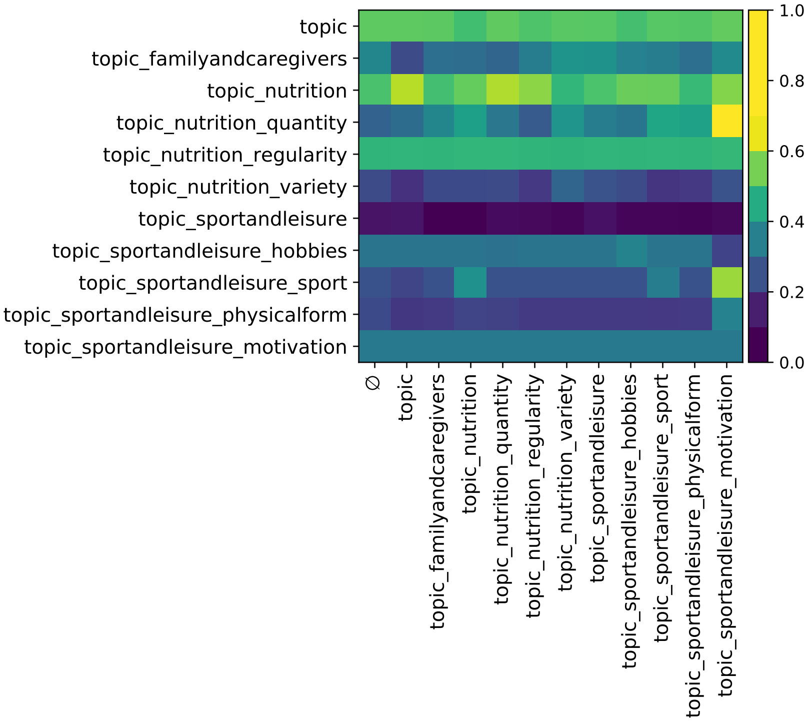

Figure 7: F1 score matrix for every combination of task and transfer in the Empathic dataset.

tences it contains. This is the largest task in terms of number of sentences after the root task, which is heavily unbalanced towards the topic_other label. This imbalance makes the transfer less useful to other topics, due to the characteristics of the benefited label. The topic_other label contains all the sentences that occur during a conversation and that do not belong to any topic, such as greetings, backchannel information, agreements, disagreements and others. All this information that is absorbed by the trained model, and transferred later on, does not seem to be helpful at all. Nevertheless, we can find some tasks with remarkable improvements such as topic_nutrition or topic_nutrition_variety, which can examined in detail in Figures 9 and 10 respectively. In addition, every task except topic_sportandleisure_motivation exhibited some improvement due to the transfer from some models.

In addition to the analysis of the influence of the transfers in terms of performance, we have conducted an analysis on how the transfers affect the number of training epochs that the models needed to train. In the case of the WOS models, the transfers can make significant reductions as can be seen in Figure 11. This figure is similar to the F1-score figures, but, instead of performance, it illustrates the average amount of epochs that were needed to train each particular task.

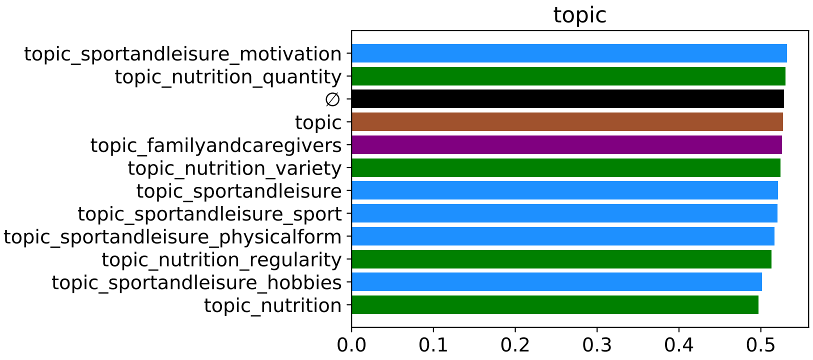

Figure 8: F1 score results for the Empathic main topic task. The blue bars belong to the sport and leisure branch, the purple bar belongs to the family and caregivers branch, and the nutrition branch is represented in green.

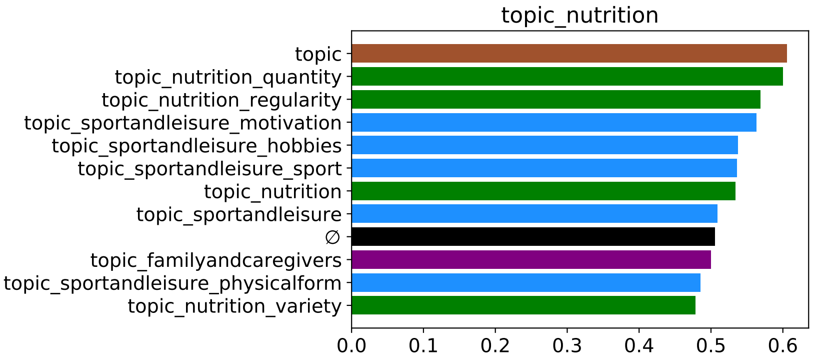

Figure 9: F1 score results for the Empathic topic_nutrition task. The blue bars belong to the sport and leisure branch, the purple bar belongs to the family and caregivers branch, and the nutrition branch is represented in green.

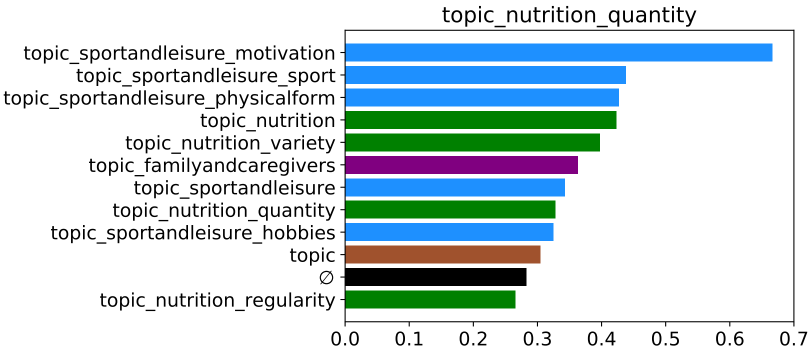

Figure 10: F1 score results for the Empathic topic_nutrition_quantity task. The blue bars belong to the sport and leisure branch, the purple bar belongs to the family and caregivers branch, and the nutrition branch is represented in green.

The figure shows an important reduction in the diagonal, but this is expected since those results represent the models that are trained with the transfer from the same task. Another pattern that can be spotted is the influence of the *Topic* transfer, which reduces the training epochs in 4 out of 6 tasks. Nevertheless, for each task there are particular cases that make the epochs needed reduce more drastically. Unfortunately, these reductions are not related to the increases in performance, which leads us to deduce that some transfers lead the training process to undesired local minima.

In the case of the Empathic project models, the results are illustrated in Figure 12. The patterns found are very similar to the ones found in the WOS dataset. The diagonal represents a reduction in every case, and the topic transfer reduces the number of epochs in 7 out of 10 transfers. As can be seen in the WOS epochs results, there is no apparent relation between the epochs needed and the performance of the models.

Figures 13 and 14 present a summary of the effectiveness of the transfer strategies we have used in the paper. In these figures, the F1 score value for each task without transfer learning is compared with the best result with transfer learning for each task. It can be appreciated that almost every task could be improved by means of the transfer from another task. Unfortunately, it is not always guaranteed that transfers from more general nodes will benefit the more specific ones. This indicates the existence of other factor that influence the outcome of the transfer, for instance, how balanced the classification tasks are or how related the topics are. These factors deserve further investigation.

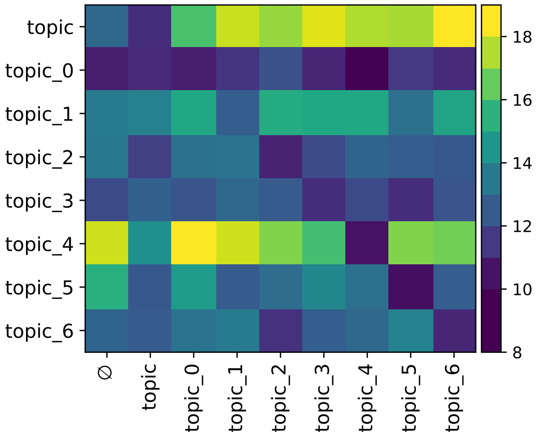

Figure 11: Number of epochs needed for training for every combination of task and transfer in the WOS dataset.

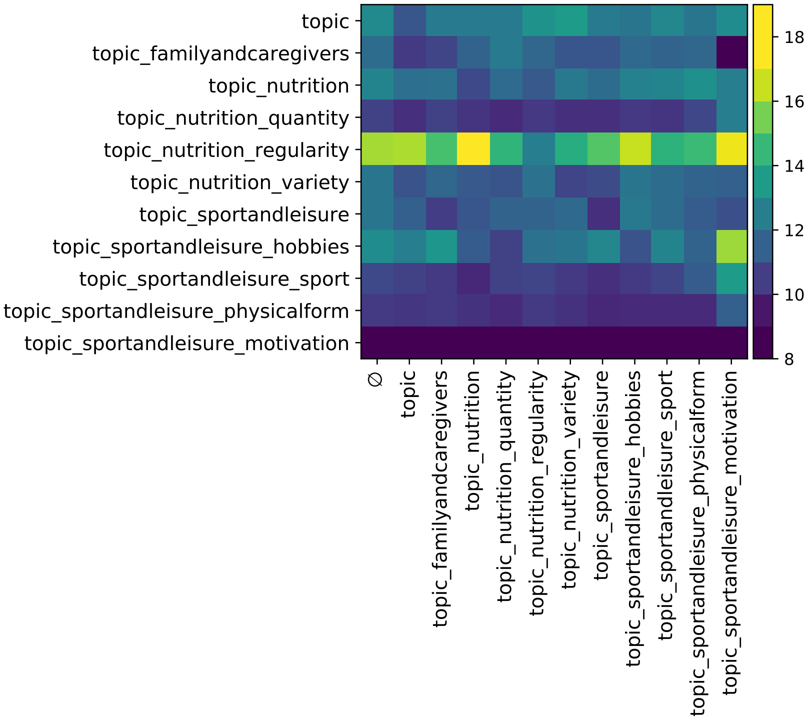

Figure 12: Number of epochs needed for training for every combination of task and transfer in the Empathic dataset.

7 Conclusions

While machine learning methods, and particularly deep neural networks, are increasingly applied to natural language processing tasks such as topic classification, the performance of these algorithms strongly depends on the availability of data. In some particular tasks such as hierarchical classification, the lack of data can produce a more critical effect. In this paper we have proposed knowledge transfer for a hierarchical topic classification problem for which labeled data is scarce.

We have proposed different ways of transferring knowledge between the hierarchical classification problems, taking into consideration the hierarchical structure of the problem as a way to research the relationship between tasks in the hierarchy and the outcome of the transfer. While our results show that transfer learning in hierarchical topic classification is a useful tool to improve the performance of the models for inner nodes, extracting or defining a procedure to predict when the transfer will be successful has proved to be an elusive goal, at least for the two datasets considered in our study.

Regarding training time, transfer learning can help to reduce the number of epochs needed to train a particular task. In general, the transfers made from a more general task help to reduce this training time. Nevertheless, the best performance and the best training time are not related.

This transfer learning method does not substitute other methods for the same goal, but complements them. Therefore it should be considered as part of the repertoire of methods for any hierarchical problem where the lack of data is a relevant issue.

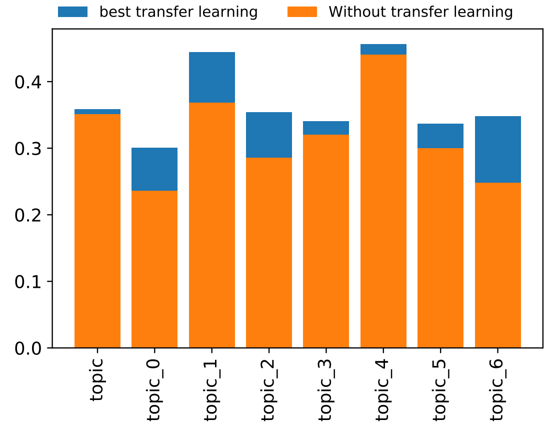

Figure 13: F1 score values of the models for each task without transfer learning compared to their best transfer learning result.

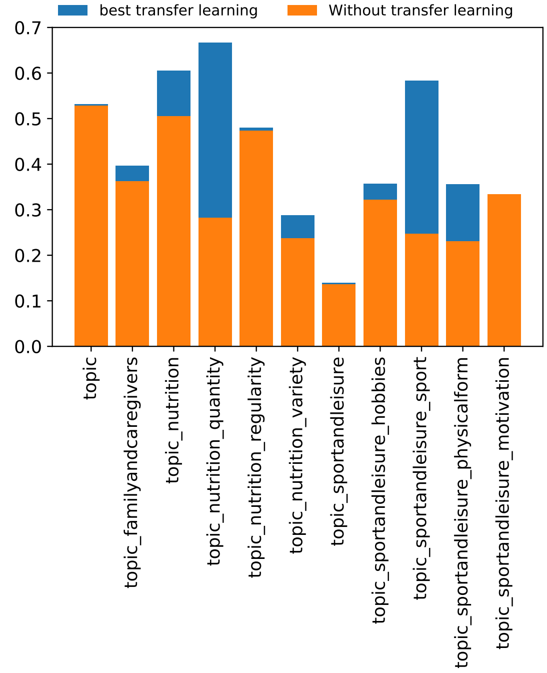

Figure 14: F1 score values of the models for each task without transfer learning compared to their best transfer learning result.

8 Acknowledgments

The research presented in this paper is conducted as part of the project EMPATHIC that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769872. Jose A. Lozano and R. Santana acknowledge support by the Spanish Ministry of Science, Innovation and Universities (Project TIN2016-78365-R), and the Basque Government (IT1244-19 and ELKARTEK Programs).

References

[1] J. Allen and M. Core. Draft of DAMSL: Dialog act markup in several layers, 1997.

[2] H. Bunt. The DIT++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24, 2009.

[3] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[4] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*, 2016.

[5] R. Johnson and T. Zhang. Supervised and semi-supervised text categorization using LSTM for region embeddings. *arXiv preprint arXiv:1602.02373*, 2016.

[6] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 364–371. IEEE, 2017.

[7] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes. Web of science dataset. *DOI: https://doi.org/10.17632/9rw3vkcfy4*, 6, 2018.

[8] J. Y. Lee, F. Dernoncourt, and P. Szolovits. Transfer learning for named-entity recognition with neural networks. *arXiv preprint arXiv:1705.06273*, 2017.

[9] M. M. Lopez and J. Kalita. Deep learning applied to NLP. *arXiv preprint arXiv:1703.03091*, 2017.

[10] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang. Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 80:14 – 23, 2015.

[11] Y. Meng, J. Shen, C. Zhang, and J. Han. Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6826–6833, 2019.

[12] C. Montenegro, A. López Zorrilla, J. Mikel Olaso, R. Santana, R. Justo, J. A. Lozano, and M. I. Torres. A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly. *Multimodal Technologies and Interaction*, 3(3):52, 2019.

[13] C. Montenegro, R. Santana, and J. A. Lozano. Data generation approaches for topic classification in multilingual spoken dialog system. In *Proceedings of the 12th Conference on PErvasive Technologies Related to Assistive Environments Conference (PETRA-19)*, pages 211–217. ACM, 2019.

[14] H. Morioka, A. Kanemura, J.-I. Hirayama, M. Shikauchi, T. Ogawa, S. Ikeda, M. Kawanabe, and S. Ishii. Learning a common dictionary for subject-transfer decoding with resting calibration. *Neuroimage*, 111:167–178, 2015.

[15] S. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[16] S. Pareti and T. Lando. Dialog intent structure: A hierarchical schema of linked dialog acts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pages 2907–2914, 2018.

[17] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber. The Penn Discourse Treebank 2.0. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pages 2961–2968, 2008.

[18] R. Raina, A. Y. Ng, and D. Koller. Constructing informative priors using transfer learning. In *Proceedings of the 23rd International Conference on Machine Learning ICML-2006*, pages 713–720, New York, NY, USA, 2006. ACM Press.

[19] R. Santana, L. Marti, and M. Zhang. Gp-based methods for domain adaptation: using brain decoding across subjects as a test-case. *Genetic Programming and Evolvable Machines*, 20(3):385–411, 2019.

[20] J. Schmidhuber and S. Hochreiter. Long short-term memory. *Neural Comput*, 9(8):1735–1780, 1997.

[21] C. N. Silla and A. A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, 2011.

[22] K. Sinha, Y. Dong, J. C. K. Cheung, and D. Ruths. A hierarchical neural attention-based text classifier. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 817–823, 2018.

[23] M. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *The Journal of Machine Learning Research*, 10:1633–1685, 2009.

[24] M. L. Torres, J. M. Olaso, C. Montenegro, R. Santana, A. Vazquez, R. Justo, J. A. Lozano, S. Schloegl, G. Chollet, N. Dugan, M. Irvine, N. Glackin, C. Pickard, A. Esposito, G. Cordasco, A. Troncone, D. Petrovska-Delacretaz, A. Mtibaa, M. A. Hmani, M. S. Korsnes, L. J. Martinussen, S. Escalera, C. Palmero-Cantarino, O. Deroo, O. Gordeeva, J. Tenerio-Laranga, E. Gonzalez-Fraile, B. Fernandez-Ruanova, and A. Gonzalez-Pinto. The EMPATHIC Project: Mid-term Achievements. In *Proceedings of the 12th Conference on PErvasive Technologies Related to Assistive Environments Conference (PETRA-19)*, pages 629–638. ACM, 2019.

[25] X. Wei, H. Lin, L. Yang, and Y. Yu. A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. *Information*, 8(3):92, 2017.

[26] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, 2016.