Few-Shot Emotion Recognition in Conversation with Sequential Prototypical Networks

Gaël Guibon[1,2], Matthieu Labeau[1], Hélène Flamein[2], Luce Lefeuvre[2], and Chloé Clavel[1]

[1]LTCI, Télécom-Paris, Institut Polytechnique de Paris
[2]Direction Innovation & Recherche SNCF
{gael.guibon,matthieu.labeau,chloe.clavel}@telecom-paris.fr
{ext.gael.guibon,helene.flamein,luce.lefeuvre}@sncf.fr

Abstract

Several recent studies on dyadic human-human interactions have been done on conversations without specific business objectives. However, many companies might benefit from studies dedicated to more precise environments such as after sales services or customer satisfaction surveys. In this work, we place ourselves in the scope of a live chat customer service in which we want to detect emotions and their evolution in the conversation flow. This context leads to multiple challenges that range from exploiting restricted, small and mostly unlabeled datasets to finding and adapting methods for such context. We tackle these challenges by using Few-Shot Learning while making the hypothesis it can serve conversational emotion classification for different languages and sparse labels. We contribute by proposing a variation of Prototypical Networks for sequence labeling in conversation that we name ProtoSeq. We test this method on two datasets with different languages: daily conversations in English and customer service chat conversations in French. When applied to emotion classification in conversations, our method proved to be competitive even when compared to other ones. The code for ProtoSeq is available at `https://github.com/gguibon/ProtoSeq`.

1 Introduction

There has been a recent surge in research focusing on analyzing dyadic human to human interactions. Many of these studies (Poria et al., 2017; Zadeh et al., 2018a,b; Majumder et al., 2019) focus on emotion recognition in conversations (ERC) taking into account multiple data modalities. Moreover, most of the progress made in ERC has been done without factoring in constraints corresponding to specific but prominent industrial applications, like customer service. This is partly due to studies focusing on using artificial datasets (Li et al., 2017; Busso et al., 2008) made of mock-up conversations to facilitate result replication and comparison. A few existing studies address customer service applications (Mundra et al., 2017; Yom-Tov et al., 2018; Maslowski et al., 2017) and show the difficulties to deal with such in-the-wild and domain-specific data.

In this work, we focus on data from a live chat support in which we want to detect emotions and their evolution in the conversational flow. This setting corresponds to a human dyadic conversation, albeit with a specific business-related objective. We make the hypothesis that the emotion flows of the visitor and the operator will bring information on the quality of the service and help operators better assist customers. This hypothesis is close to relevant studies on the importance of emotions and empathy in dyadic call center conversations (Alam, 2017; Alam et al., 2018). This specific setting leads to multiple challenges: indeed, it is difficult and costly to label this kind of data — and even then, these exchanges are very sparse in emotions, most of the labels associated with utterances being neutral. To maximize data efficiency, we use Few-Shot Learning (FSL), and adapt a popular approach to our highly unbalanced data. By setting up this approach in an episodic fashion (Ravi and Larochelle, 2016), we join studies on ERC and studies on FSL to tackle this industrial use-case.

We contribute by proposing a variant to Prototypical Networks (Snell et al., 2017) dedicated to ERC on data produced by company services, framing it as a sequence labeling task. We modify the original model by allowing it to consider the whole conversational context when making predictions, through a sequential context encoder and the use of Conditional Random Fields (CRF) on top of the model. We test our method on two datasets, in two different languages. The first one, made of daily conversations in English, allows us to compare ourselves to previous methods, while the second one, made of private data from a live chat customer ser-

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6858–6870
November 7–11, 2021. ©2021 Association for Computational Linguistics

vice, allows us to conduct a performance analysis in our target setting. We also present the latter dataset, along with its annotation process.

This paper is organized as follows. First, we sum up the related work on textual ERC and FSL in conversations (Section 2). Then we present the datasets along with the emotional annotation scheme and the annotation campaign set up for the customer service live chats dataset (Section 3). We continue by thoroughly presenting the Sequential Prototypical Networks (Section 4) before looking at the achieved results on both datasets (Section 5). Finally, we present the limitations of such a system (Section 6) and conclude (Section 7).

2 Related Work

Emotion Recognition in Conversations In recent years, the widening scope of emotion detection tasks led to the rise of another sub-topic: detecting emotions in conversations. This research topic, commonly referred to as ERC, gained popularity when Poria et al. (2017) first applied recurrent neural networks (RNN) (Jordan, 1997) to multi-modal emotion recognition in conversations. This led to many improvements (Zadeh et al., 2018a,b; Hazarika et al., 2018; Majumder et al., 2019). Among those, Majumder et al. (2019) used 3 Gated Recurrent Units (GRU) (Cho et al., 2014) units, one for each context representation target (speaker, utterance, emotion). Studies on ERC applied to text followed, mainly built on an artificial conversation dataset named DailyDialog (Li et al., 2017). (Zhong et al., 2019) incorporated a knowledge base into the network using context-aware attention and hierarchical self-attention using Transformers (Vaswani et al., 2017). Ghosal et al. (2019) uses graph neural networks to deal with context propagation limitations. These approaches in ERC consider the conversational context surrounding the current utterance; on the other hand, some recent studies consider it as a sequence and tackled ERC through a sequence labeling task (Wang et al., 2020). We follow this last approach and consider the ERC task as a sequence labeling task. However, these supervised approaches are difficult to use, as it is hard to find a sufficient amount of conversations labeled with emotions. Hence, in this paper, we approach ERC as a few-shot learning problem.

Few-Shot Learning FSL (Miller et al., 2000; Fei-Fei et al., 2006; Lake, 2015) is suitable to tackle this data limitation. It aims at generalizing faster, leading to a lower dependency on data quantity. It is mainly set up through episodic composition (Ravi and Larochelle, 2016) which recreates the few-shot learning setting by working with small training episodes. Several learning methods are based on metric-learning: Siamese Networks, which share some weights, are used to learn a metric between examples (Koch et al., 2015). Matching Networks (Vinyals et al., 2017) use the training examples to find the weighted nearest neighbors (Vinyals et al., 2017). Prototypical Networks Snell et al. (2017) consider averaged class representations from the training examples and a cosine distance to compare the elements to these class representations. Relation networks replace the Euclidean by the deep neural network which aims at training a distance metric (Sung et al., 2018).

In this work, we consider approaches based on Prototypical Networks. As Al-Shedivat et al. (2021) recently showed it, such approaches are the most efficient when working with a low amount of training samples. Many variants have been proposed, on different tasks and topics such as relation classification in text (Gao et al., 2019; Hui et al., 2020; Ren et al., 2020), sentiment classification in Amazon comments (Bao et al., 2020), named entity recognition (Fritzler et al., 2019; Hou et al., 2020; Perl et al., 2020; Safranchik et al., 2020), or even speech classification in conversation (Koluguri et al., 2020). This surge of interest on applying few-shot learning to these topics can be attributed to specific datasets, such as Few-Rel (Han et al., 2018) for relation classification. While ERC is mainly considered in a fully supervised learning setting, we intend to view it as a few-shot learning sequence labeling class. In this paper, we propose the first few-shot learning approach on ERC using sequence labeling through adapting Prototypical Networks. We compare our method to the original Prototypical Networks (Snell et al., 2017) and to a variant dedicated to named entity recognition (Fritzler et al., 2019) that is easily applicable to our task.

3 Data

To be able to both study the behavior of our model in its targeted industrial use-case, and allow performance comparison with baselines, we will work with two very different corpora: our proprietary live chat customer service dataset, and DailyDialog (Li et al., 2017). In both datasets, messages are labeled with emotions while considering the context of the conversation. However, they vary considerably in their topics and lexical fields: ordinary matters for DailyDialog and railway related customer service for the live chats. They also vary in the assumptions they make about the speakers : while the topics discussed in DailyDialog imply a sense of proximity, the live chat customer service involve complete strangers with pre-existing emotional states (*e.g.* the visitor is already stressed due to a refund issue). Both datasets' statistics can be found in Table 1.

3.1 DailyDialog

DailyDialog is a dyadic conversation dataset in English whose purpose is to represent casual, everyday interactions between people, in order to facilitate training and sharing of dialog systems. The exchanges in DailyDialog are artificial conversations which are neither dedicated to a specific topic nor task-oriented: they mainly deal with relationships, everyday life, and work. Each utterance corresponds to a speaker turn, and is labeled with one of 7 labels: the 6 basic emotions (anger, disgust, fear, joy, sadness, and surprise) and "no_emotion" denoting the absence of one. The "no_emotion" label represents 80% of the corpus, leading to a very unbalanced dataset with an average length of 8 messages per conversation and a maximum of 35 messages. For this dataset, the inter annotator agreement achieved 78.9%. We choose DailyDialog for comparison and reproducibility purposes, as it is often used for ERC. In this work, we use the train/val/test splits provided by (Zhong et al., 2019).

3.2 Live chat customer service

Our primary objective is to detect emotions in conversations from a customer service live chat involving a visitor (*i.e.* the customer looking for help) and an operator (*i.e.* an employee being there to assist the visitor and better satisfy him). The corpus is written in French and is made of 5,000 conversations from which we annotate a subset of 1,500 conversations, leading to a total of 20,754 messages. The average message length is higher than DailyDialog, with 15.14 messages per conversation. We do not have a way to identify real speaker turns. Indeed, a speaker turn is not necessarily the sequence of contiguous segments corresponding to a same speaker because there could be a time delay between two messages of a same speaker, indicating that the speaker is changing the topic. Because all our messages have a very short time difference we prefer not to automatically infer speaker turns and consider the message as the unit of analysis. This means the conversation context is a sequence of messages instead of a sequence of speaker turns which could have contained one or more messages artificially glued together.

Dataset	DD	Chat
Language	English	French
Type	Artificial	Customer Service
Max Msg/Conv	35	84
Avg Msg/Conv	8	13
Labels	7	11
Labels for eval	6	9
Nb. Conv.	13,118	1,500

Table 1: Statistics for both datasets DailyDialog (DD) and Live Chat Customer Service (chat).

Two annotators were involved in the process, which unrolled as follows: first, each message is labeled with an emotion. Once all the messages in a conversation have been assigned an emotion label, the conversation is labeled with a visitor satisfaction score (ranging from -3 to 3), and the status of the customer request ("solved", "test_required", "out of scope", or "aborted"). After a preliminary study of the corpus, we identify 10 emotion labels as relevant in this corpus: neutral, surprise, amusement, satisfaction[1], relief, fear-anxiety-stress, sadness, disappointment, anger, and frustration. Compared to (Chowdhury et al., 2016), we consider the satisfaction at the conversation level and we are more precise with not only positive, neutral, and negative levels, but also with 4 additional intermediate levels (from -3 to +3 included). We have also a higher number of emotions, with 10 emotions instead of 4, with more precise emotions such as relief for instance. In our customer service interface, some alerts are automatically prompted for specific actions such as “user x left the chat” or “operator sent a link”. We call these “alerts”, and they are labeled as “no_emotion”. The “neutral” label means that the emotional content of the message, written by a human, has been considered as neutral by the annotator. Figure 1 illustrates the distribution of emotion labels in the Live Chat Customer Service

[1] It is interesting to notice here that in the current application setting, "joy" label has been replaced by "satisfaction", because it is more suited to the customer relationship context (Danesi and Clavel, 2010).

dataset. We can see that the neutral label is the most frequent by a large margin.

The Cohen's κ scores obtained on the 3 label types correspond to substantial agreement at the message level and moderate agreement at the conversation level (Landis and Koch, 1977). κ-score is given for 3 label types: 1) the emotions at the message level ($\kappa = 0.65$); 2) the visitor's satisfaction at the conversation level ($\kappa = 0.45$); and 3) the request's status at the conversation level ($\kappa = 0.46$). Similarly to DailyDialog, the "neutral" label represents 81.5% of the corpus, resulting in another very unbalanced dataset in terms of emotions, as rendered obvious by Figure 1. Excluding this label gives a slightly more balanced label set, as the satisfaction represents 44.9% of the other emotions, and the "frustration" 20.8%.

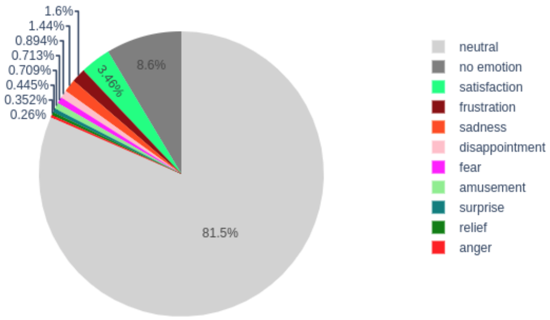

Figure 1: Emotion Distribution in Live Chat Customer Service

To tackle our hypothesis that the conversational emotion flow can define the overall visitor satisfaction, we calculate the Pearson correlation between the emotions at the message level and the global satisfaction of the visitor at the conversation level. These scores show the more extreme the emotion, the greater the correlation with the satisfaction score is[2].

4 Methodology

Formally, our dataset $\mathcal{D}$ is comprised of conversations $(C_1, C_2, \ldots, C_{|D|})$, which are in turn made of utterances: $C_i = (u_1, u_2, \ldots, u_{|C_i|})$. To each of these utterances is associated an emotion label, giving a sequence of labels by conversation: $Y_i = (y_1, y_2, \ldots, y_{|C_i|})$. Finally, an utterance is a sequence of words, $u_j = (w_1^j, w_2^j, \ldots, w_{|u_j|}^j)$.

Emotion	κ-score
Amusement	0.1115
Anger	0.1608
Disappointment	0.1609
Frustration	0.1193
Neutral	0.3187
Fear	0.1111
Satisfaction	0.2068
Relief	0.1429
Surprise	0.1885
Sadness	0.2860
Global	0.6499
Global w/o Neutral and no_emotion	0.3885

Table 2: By-category agreement scores for emotions in Live Chat Customer Service

4.1 Episodic learning

We use the episodic approach (Ravi and Larochelle, 2016), which simulates a context where only a few examples per class are available during training and the model must adapt during testing. This approach perfectly fits into our need for FSL. The episodic composition is defined by setting the number of classes (ways) N_C, the number of examples per class N_S (shots) and the number of elements to label N_Q (queries). In our experiments with DailyDialog, the task is 5-shot 7-way 10-query, and when using our customer service chats, the number of classes changes, making it a 5-shot 11-way 10-query. In the context of sequential ERC, this means that for each episode we train the model on 5 conversations (*i.e.* sequences to label) per emotion and apply it to 10 conversations per emotion. We identify a sequence as belonging to the target class set if at least one message is labeled with the target class in the sequence. This means that the number of example messages in each support set S_k of class k can vary (with a minimum of N_S elements), while the number of sequences is fixed.

4.2 ProtoSeq: Prototypical Networks for Emotion Sequence Labeling

In order to apply FSL to ERC, we choose to base our model on Prototypical Networks (Snell et al., 2017), which create prototypes from the average of the embeddings of the words forming the utterance. Our proposed model, ProtoSeq, builds on this by factoring in conversational context and performing sequence labeling, thus allowing the use of both input and output dependencies when applying

[2] See appendix A

FSL to ERC. ProtoSeq is divided into four main components, applied at each consecutive level of granularity of the data.

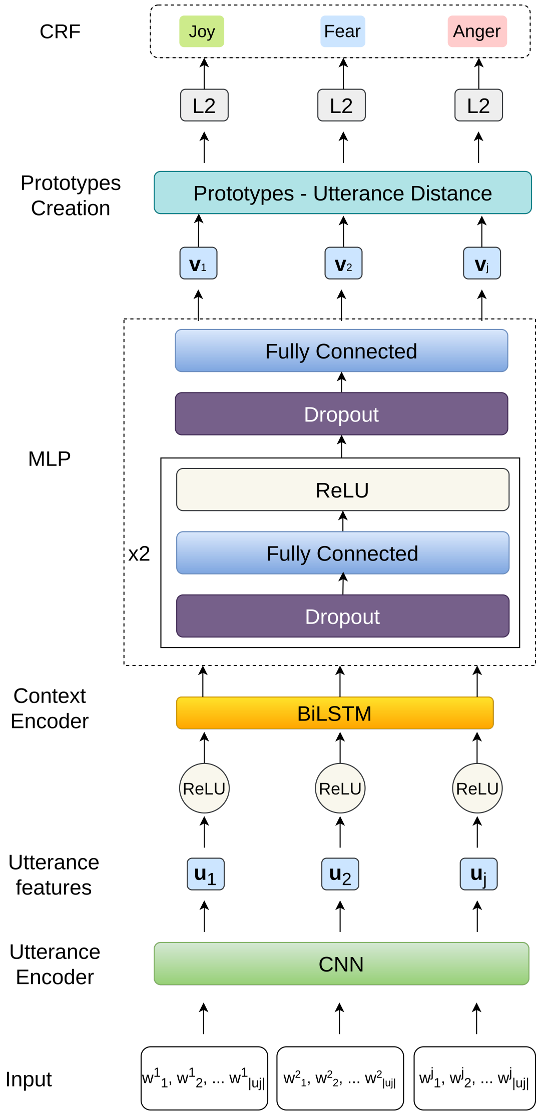

Figure 2: ProtoSeq Global View

Utterance Encoder: Similar to the encoder of the Prototypical Networks, our utterance encoder f_u reduces the utterance u_i to only one vector:

$$\mathbf{u}_j = f_u(w_1^j, w_2^j, \ldots, w_{|u_j|}^j)$$

The architecture of our encoder is based on the Convolutional Neural Network (CNN) described by (Kim, 2014) , which makes tokens through different convolution filters and merges the representation through max-over-time pooling.

Context Encoder: After applying a non-linear activation (ReLU), we use a Bi-directional LSTM layer (BiLSTM) (Huang et al., 2015), to integrate contextual information from the conversation, thus following the trend initiated by (Poria et al., 2017) in ERC to use a recurrent context encoder. We obtain contextual utterance representations $\mathbf{v}_j$:

$$\mathbf{v}_j = f_c([\mathbf{u}_k]_{k=1}^{j-1}, \mathbf{u}_j, [\mathbf{u}_k]_{k=j+1}^{|C_i|})$$

As we work in a few-shot learning setting, we try not to over-complexify our model, hence we do not add a transformer-based global context encoder (Wang et al., 2020) on top of the BiLSTM.

Prototypes Creation: We feed the output of the context encoder to a multi-layer perceptron made of 2 fully connected layers with dropout and ReLU. The resulting representations are then used to create prototypes $\mathbf{c}$: for the class k,

$$\mathbf{c}_k \leftarrow \frac{1}{N_\mathcal{C}} \sum_{(u_j, y_j) \text{ with } y_j = k} MLP(\mathbf{v}_j)$$

where $N_\mathcal{C}$ is the number of classes, and MLP refers to Multi-Layer Perceptron.

Sequence Prediction: We compute the euclidean distance from the contextual representation of the utterance to each class prototype. The predicted label $\hat{y}_j$ to each utterance u_j is the class corresponding to its closest prototype:

$$\hat{y}_j \leftarrow \arg\min_{k \in \mathcal{C}} d(MLP(\mathbf{v}_j), \mathbf{c}_k)$$

We allow our model to consider dependencies between the labels, we add a final CRF layer on top of label prediction, the emission scores being the euclidean distances for each utterance. Overall, our model is a variation of the traditional BiLSM-CRF model, based on prototypical networks.

4.3 Experimental protocol

We follow the setting used by (Bao et al., 2020) by considering a training epoch as a set of 100 random episodes from the training set, and applying a validation step made of 100 random episodes from the validation set after each epoch. We test our model using 1,000 random episodes from the test set. The maximum number of epochs is set to 1,000, but when the F1-micro score does not improve for 100 consecutive epochs, we stop the training and reload the best model's weights. We use the Adam (Kingma and Ba, 2017; Loshchilov and Hutter, 2019) optimizer to train the model while maximizing the log-likelihood loss of the correct emotion sequences in the query set Q_k

$$\mathcal{L} = \sum_{C \in Q_k} \sum_{j=1}^{|C|} \log(p(\hat{y}_j \mid u_j, C))$$

During inference, we apply the Viterbi algorithm to output the best-scoring sequence of labels. We do not cut on either utterance or conversation length. To obtain an initial token representation, we use pre-trained FastText (Bojanowski et al., 2017) embeddings from Wiki News[3] for English (DailyDialog), and from Common Crawls[4] for French (customer service live chats). Both sets of embeddings are of dimension 300 and both datasets are tokenized with NLTK[5]. We choose our hyper-parameters using a very targeted grid search for the learning rate (set to $1e3$ for all the experiments) and manual tuning for the other parameters. In the following, we experiment with several variants of our model, each having dedicated hyper-parameters.

- **ProtoSeq**: We use hyper-parameters from Kim (2014) for the CNN: 50 filters with windows 3 different sizes (3, 4 and 5). We use one BiLSTM layer with 150 hidden units in order to fit to the 300 dimensions of the inputs considering the two directions.

- **ProtoSeq-CNN**: A lighter version of our model, without the BiLSTM context-encoder. The CNN configuration follows the same parameters from Kim (2014).

- **ProtoSeq-Tr**: A ProtoSeq with a 2-layers Transformer-based utterance encoder with 4 attention heads and a hidden size of 300. The global dropout is set to 0.2 while the position encoder dropout is set to 0.1.

- **ProtoSeq-AVG**: A ProtoSeq where the utterance encoder is just an average of the token representations. However, it should be noted that the averaging process excludes the padding elements in the utterances.

5 Results

Tables 3 and 4 show the performance of the model using the micro F1-score. We use the protocol usually followed by the literature and do not take into account the majority class "no_emotion" as it represents 80% of the DailyDialog corpus. This allows performance comparison with related work on ERC through supervised learning. We do the same for the Live Chat Customer Service corpus by ignoring the "neutral" label.

Comparison to supervised learning DailyDialog is used to compare our FSL approach with recent supervised learning results on ERC. As expected, our best FSL model, ProtoSeq, yields lesser performance than supervised approaches. The latter presuppose the availability of a sufficiently large amount of annotated data and their performance thus represents the upper bound of the expected results. More precisely, we focus on the difference between ProtoSeq with a state-of-the-art supervised model, CESTa (Wang et al., 2020), which is computation-heavy. Indeed, CESTa is a contextualized emotion sequence tagging model which considers the fusion of a combination of a transformer and BiLSTM as the global context encoder with a recurrent individual context encoder before feeding a CRF layer. CESTa achieves 63% in micro F1-score in a fully supervised learning approach. ProtoSeq, much lighter, achieves a 31% micro F1 score, demonstrating the potential of FSL for sequence labeling when available data is scarce, especially when many supervised approaches obtained F1-scores around 50%. While using the Live Chat Customer Service dataset, we only change the initial embeddings from English to French, and apply the two best models according to 3: CESTa and KET (Zhong et al., 2019). The CESTa implementation yielded inconclusive results[6], this is why we present the KET results on our specific corpus in Table 4. KET relies on ConceptNet (Speer et al., 2017), a multilingual knowledge base. Thus, we only switch from GloVe embeddings (Pennington et al., 2014) to French FastText ones in order to ensure comparison with our ProtoSeq model. As expected, performance is lower on the Live Chat Customer Service corpus.

Few-shot learning baselines We consider two baselines. We apply the original Prototypical Networks (Snell et al., 2017), only retrieving the labels using the euclidean distance to class prototypes. We also apply the WarmProto-CRF (Frit-

[3] https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M.vec.zip

[4] https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.fr.300.bin.gz

[5] https://www.nltk.org/

[6] We present in our code an implementation of CESTa following the paper's descriptions. On our dataset, it only labeled the two majority classes 'no_emotion' and 'neutral', leading to a null F1 (micro).

Model	F1 (weighted)	MCC	F1 (micro)
Supervised Learning			
cLSTM			0.4990
CNN (Kim, 2014)			0.4934
CNN+cLSTM (Poria et al., 2017)			0.5184
BERT BASE (Devlin et al., 2019)			0.5312
DialogueRNN (Majumder et al., 2019)			0.5164
KET (Zhong et al., 2019)			0.5337
CESTa (Wang et al., 2020)			**0.6312**
Few-Shot Learning			
Proto (Snell et al., 2017)	0.2377 ±0.0136	0.3448 ±0.0105	0.2141 ±0.0141
WarmProto-CRF (Fritzler et al., 2019)	0.2384 ±0.0383	0.3403 ±0.0365	0.2607 ±0.0381
ProtoSeq-AVG	0.1312 ±0.0201	0.2622 ±0.0225	0.1643 ±0.0258
ProtoSeq-Tr	0.1694 ±0.0293	0.3329 ±0.0241	0.2557 ±0.0317
ProtoSeq-CNN	0.2244 ±0.0359	0.3494 ±0.0182	0.2560 ±0.0275
ProtoSeq	0.3522 ±0.0302	0.3922 ±0.0233	**0.3181** ±0.0276

Table 3: Sequence labeling on DailyDialog splits (Zhong et al., 2019) (seq size = 35). Top section shows supervised learning results reported from related work, bottom section presents our results using few-shot learning (7 way 5 shot 10 query). MCC = multi-class Matthews Correlation Coefficient (MCC). ± = test episodes variance..

Model	F1 (weighted)	MCC	F1 (micro)
Supervised Learning			
KET (Zhong et al., 2019)			**0.4143**
Few-Shot Learning			
Proto (Snell et al., 2017)	0.1749 ±0.0481	0.3333 ±0.0133	0.1228 ±0.0194
WarmProto-CRF (Fritzler et al., 2019)	0.1556 ±0.0522	0.7212 ±0.0220	0.1918 ±0.0601
ProtoSeq-AVG	0.1297 ±0.0246	0.7163 ±0.0215	0.1582 ±0.0251
ProtoSeq-Tr	0.1774 ±0.0285	0.6695 ±0.0163	0.2208 ±0.0371
ProtoSeq-CNN	0.1197 ±0.0198	0.7336 ±0.0135	0.1581 ±0.0180
ProtoSeq	0.3022 ±0.0256	0.6396 ±0.0222	**0.2668** ±0.0270

Table 4: Few-shot learning results on Customer Service Live Chats (seq size = 18): 11-way 5-shot 10-query (padding & trim). MCC = multi-class Matthews Correlation Coefficient (MCC). ± = test episodes variance.

zler et al., 2019) which is a variant of Prototypical Networks designed for sequence labeling by integrating CRFs. We implement it without including the bias they created for the O label in the BIO sequence labeling task. This method uses a BiLSTM utterance encoder to further compute the prototypes with the euclidean distance.

Few-shot learning on DailyDialog Table 3 shows FSL results in the bottom section. All these models are trained in an episodic fashion, with the same episode constitution (5-shot 7-way 10-query). We can see the micro F1-score is really low with only 16.43%. By considering a ProtoSeq only using an utterance encoder based on CNN (ProtoSeq-CNN) or an utterance encoder based on a 2-layers 4-heads Transformer (ProtoSeq-Tr) we can see the score improve. The addition of the BiLSTM context encoder really enables the model to capture more information: these variants show the importance of integrating a context encoder in the model.

Few-shot learning on Customer Service Live Chats We also apply this approach on the Customer Service Live Chats, further motivated by the high annotation cost and the fact that supervised approaches on clean data such as DailyDialog did not achieve an acceptable score for this use case (starting from 70 % in micro F1 score). Besides, new conversations with evolving contents (e.g., due to the evolution of company services) are created everyday. As a consequence, it would render the ideally annotated training corpus obsolete at some point. This FSL prediction leads to lesser scores, but with the same hierarchy among variants. ProtoSeq, using a BiLSTM context encoder, yields again the best scores. The higher number of classes (with 11 classes including 9 emotions versus 7 classes including 6 emotions) may explain the overall lower numbers we observe here, compared to those we obtain on DailyDialog.

Artificial versus Real Data DailyDialog is an artificial corpus which follows standard, idealized conversations. We can see that ERC performance is quite sensible to the conversation length, which seems to confirm conclusions drawn in recent literature (Wang et al., 2020). Customer Service Live Chats being real use-case data, their length varies a lot, ranging from 2 to 85 messages (where conversations from DailyDialog go from 2 to 35 messages). However, ERC also seems to be impacted by the utterance textual content, as our data contains a lot of spelling mistakes, shortcuts, or slurs. More importantly, the visitor may often use several small messages rather than only one to transmit information; this flow may be interrupted by a message from the operator, making it impossible to detect the whole set of messages as an utterance. This is specific to online instant conversations where speakers do not necessarily wait for the complete message to be written or sent by the addressee. By contrast, DailyDialog is made of clean and perfect exchanges, where one waits for the other to send the answer. Here is an example with the following clean conversation subset from DailyDialog.

> **A**: Does your family have a record of your ancestors?
> **B**: Sure. My mom has been working on our family tree for years.

This conversation would often be represented as follows in real data from instant chat:

> **Operator**: Did you make the simulation using the promo code?
> **Visitor**: I did it 5 minutes ago
> **Operator**: Ok, you have to wait 30min
> **Visitor**: but as said before, I didn't finished the "simulation" because I had to pay a 10€ ticket even th
> **Visitor**: ....even though the right one is 11.5€
> **Operator**: And the code will be available again

Moreover, specific lexical fields, relevant to the customer service being provided, can also make the task more difficult for the model.

Quantifying the impact of the CRF layer Our model benefits from the addition of a final CRF layer to compute the best possible output sequence. This allows the model to generalize faster and to achieve a higher score despite the few examples. However, the prediction stability lowers, as the standard deviation across episodes shows in Table 5. The downgrade in performance while omitting the CRF layer may be due to the label dependency it emphasizes. Indeed, without the CRF, label dependency can only be inferred from the BiLSTM context encoder. The CRF layer accentuates in-episode label dependency by allowing the prediction to be further adapted to the conversation context for each query conversation.

Model	DailyDialog	CSChats
ProtoSeq-noCRF	0.2156 ±0.0105	0.1351 ±0.0144
ProtoSeq	0.3181 ±0.0276	0.2668 ±0.0270

Table 5: Micro F1-scores without and with the final CRF layer. CSChats = Customer service live chat

Emotion Predictions Tables 6 and 7 show additional information from ProtoSeq's performance on each label. These tables present averaged scores from all episodes' query sets. We can see the predictions differ a lot depending on the target label. When applied to DailyDialog, the model has no difficulty in detecting the absence of emotion. This is to be expected as this label mainly represents the conversations. However, the prediction scores for emotion labels are imbalanced, with recall scores higher than precision on both datasets.

On DailyDialog, the anger and the sadness labels really hinders the overall prediction. However, on the Customer Service Live Chats, Table 7 shows really poor prediction for the disappointment (translated from the French "déception" label) and fear labels. Actually, in this dataset the precision seems to be the main issue with only the frustration and satisfaction labels being somewhat correctly labeled. Given the model and the task, the detailed results obtained on both datasets show that performance score may benefit from the usage of macro F1-score along with the micro F1-score. Indeed, be it DailyDialog or Customer Service Live Chats, the multi-class prediction of sequence tagging is really sparse, and thus leads to imbalanced prediction, even while using an episodic strategy.

Moreover, the gap between results on DailyDialog and the ones on the Customer Service Live Chats confirms the necessity for the ERC-related studies to focus on real conversation datasets whenever it is possible.

	precision	recall	f1-score
no emotion	0.98	0.91	0.94
anger	0.24	0.38	0.30
disgust	0.12	0.29	0.17
fear	0.58	0.55	0.57
happiness	0.39	0.63	0.49
sadness	0.07	0.21	0.11
surprise	0.17	0.37	0.24

Table 6: Additional results on DailyDialog with our ProtoSeq prediction.

	precision	recall	f1-score
no emotion	1.00	1.00	1.00
surprise	0.06	0.10	0.07
amusement	0.12	0.54	0.20
satisfaction	0.47	0.60	0.53
relief	0.21	0.23	0.22
neutral	0.92	0.79	0.85
fear*	0.02	0.01	0.01
sadness	0.08	0.18	0.11
disappointment	0.03	0.07	0.04
anger	0.02	0.40	0.03
frustration	0.45	0.43	0.44

Table 7: Additional results on customer service live chats with our ProtoSeq prediction. We define the "fear" label as "fear/anxiety/stress". "no emotion" is only used for automatic chat prompts.

6 Limitations

While the ProtoSeq model seems to be suitable for FSL in ERC, it still has inherent limitations related to its architecture. ProtoSeq uses a CRF as its final layer, leading to a sequence labeling optimizer that does not take the order into account. While this yields better performance, it does not guarantee that the order information retrieved from the context encoder is wisely used, especially since we use the euclidean distances to class prototypes as emission scores for the sequence labeling. An ordered-prediction approach may allow the model to better assist operators in real-time during their decision process.

Another limitation of our model is that it may be difficult to adapt to changes in the context in which customer service is provided. Indeed, the type of service or the plaftorm used may lead to lexical field changes or very different emotional states for the incoming visitors.

7 Conclusion

In this paper, we presented the first study on emotion recognition in conversations using few-shot learning. We proposed a variant of Prototypical Networks taking into account the emotion recognition as a sequence labeling task while allowing fast convergence. When compared to other prototypical networks for sequence labeling in few-shot, our model obtained higher scores on both DailyDialog and Customer Service Live chats. Through this work, we showed that few-shot learning is possible for this task even though it is still difficult to achieve the same performance as supervised learning approaches. This study also shows the challenges that remain when tackling in-the-wild data collected in the context of a real application.

Future work will be dedicated to the improvement of the current few-shot ERC approach by adding unlabeled elements in the support set and by investigating the addition of external business knowledge to such an approach.

Acknowledgements

This project has received funding from SNCF, the French National Research Agency's grant ANR-17-MAOI and the DSAIDIS chair at Télécom-Paris.

References

Maruan Al-Shedivat, Liam Li, Eric Xing, and Ameet Talwalkar. 2021. On data efficiency of meta-learning. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1369–1377. PMLR.

Firoj Alam. 2017. *Computational Models for Analyzing Affective Behavior and Personality from Speech and Text*. Ph.D. thesis, DIT - University of Trento.

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language*, 50:40–61.

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *International Conference on Learning Representations*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Shammur Absar Chowdhury, Evgeny A Stepanov, Giuseppe Riccardi, et al. 2016. Predicting user satisfaction from turn-taking in spoken conversations. In *Interspeech*, pages 2910–2914.

Charlotte Danesi and Chloé Clavel. 2010. Impact of spontaneous speech features on business concept detection: a study of call-centre data. In *Proceedings of the 2010 international workshop on Searching spontaneous conversational speech*, pages 11–14.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611.

Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 993–1000.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6407–6414.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. *arXiv:1810.10147 [cs, stat]*. ArXiv: 1810.10147.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. *arXiv preprint arXiv:2006.05702*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Bei Hui, Liang Liu, Jia Chen, Xue Zhou, and Yuhui Nian. 2020. Few-shot relation classification by context attention-based prototypical networks with bert. *EURASIP Journal on Wireless Communications and Networking*, 2020:1–17.

Michael I Jordan. 1997. Serial order: A parallel distributed processing approach. In *Advances in psychology*, volume 121, pages 471–495. Elsevier.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese Neural Networks for One-shot Image Recognition. *ICML*, page 8.

Nithin Rao Koluguri, Manoj Kumar, So Hyun Kim, Catherine Lord, and Shrikanth Narayanan. 2020. Meta-learning for robust child-adult classification from speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8094–8098. IEEE.

Brenden Lake. 2015. LakeEtAl2015Science-startOfFewShot.pdf. *Sciences Mag*.

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. *arXiv:1710.03957 [cs]*. ArXiv: 1710.03957.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6818–6825.

Irina Maslowski, Delphine Lagarde, and Chloé Clavel. 2017. In-the-wild chatbot corpus: from opinion analysis to interaction problem detection. In *IC-NLSSP 2017*, pages 115–120.

Erik G Miller, Nicholas E Matsakis, and Paul A Viola. 2000. Learning from one example through shared densities on transforms. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 464–471. IEEE.

Shreshtha Mundra, Anirban Sen, Manjira Sinha, Sandya Mannarswamy, Sandipan Dandapat, and Shourya Roy. 2017. Fine-grained emotion detection in contact center chat utterances. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 337–349. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Tal Perl, Sriram Chaudhury, and Raja Giryes. 2020. Low resource sequence tagging using sentence reconstruction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2692–2698, Online. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. *OpenReview*.

Haopeng Ren, Yi Cai, Xiaofeng Chen, Guohua Wang, and Qing Li. 2020. A two-phase prototypical network model for incremental few-shot relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1618–1629.

Esteban Safranchik, Shiying Luo, and Stephen Bach. 2020. Weakly supervised sequence tagging from noisy rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5570–5578.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2017. Matching Networks for One Shot Learning. *arXiv:1606.04080 [cs, stat]*. ArXiv: 1606.04080.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195, 1st virtual meeting. Association for Computational Linguistics.

Galit B Yom-Tov, Shelly Ashtar, Daniel Altman, Michael Natapov, Neta Barkay, Monika Westphal, and Anat Rafaeli. 2018. Customer sentiment in web-based service interactions: Automated analyses and new insights. In *Companion Proceedings of the The Web Conference 2018*, pages 1689–1697.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. *arXiv:1909.10681 [cs]*. ArXiv: 1909.10681.

A Correlation scores

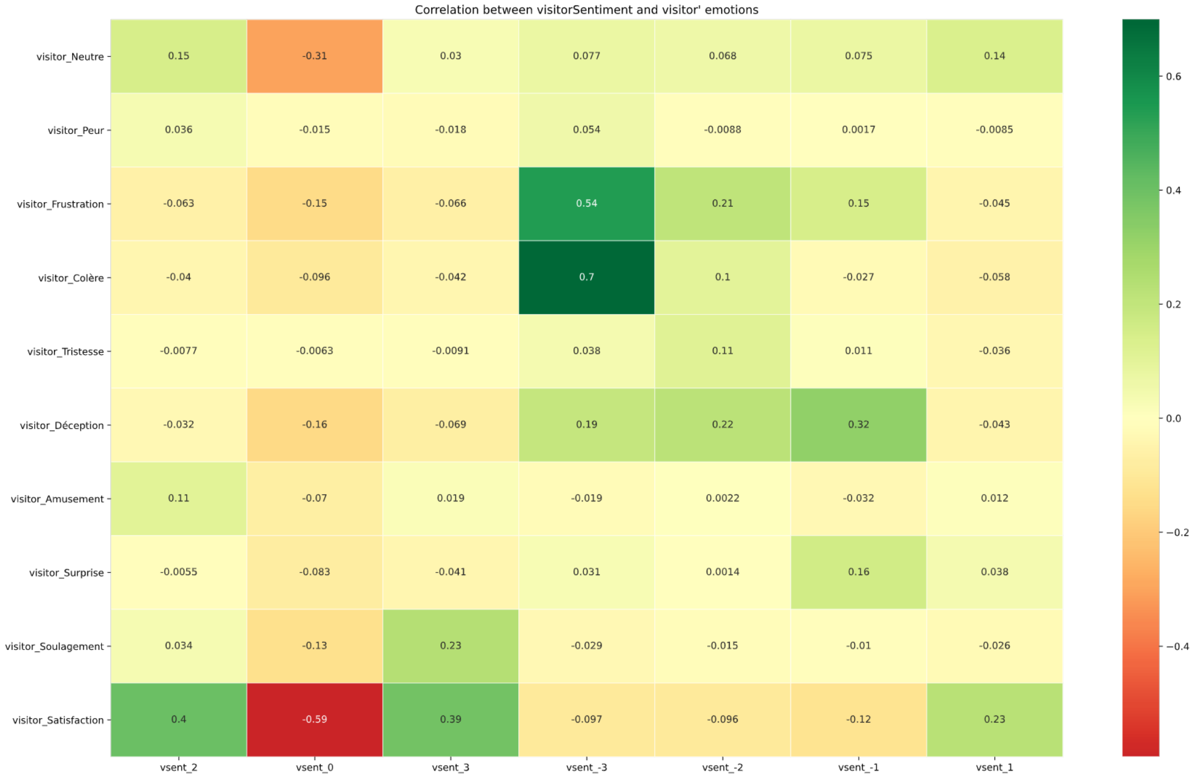

Figure 3: Pearson correlation scores between the visitor's overall satisfaction score in the conversation (vsent) and the presence of specific emotions in the messages' emotion flow (visitor_[emotion]).

Figure 3 presents the Pearson correlation scores between visitor's emotions and satisfaction for the Customer Service Live Chats. While emotions are labeled for each utterance in conversation, satisfaction is a global label for the whole conversation. This Figure shows the correlation scores are higher when the emotion is extreme within a given polarity. For instance, anger is greatly correlated to a negative satisfaction score (vsent -3) than fear or disappointment, while "Satisfaction" is more correlated to a positive overall satisfaction score (vsent +3) than "Amusement" or "Relief" are to intermediate satisfaction scores (vsent_1 or vsent_2).