Identifying users' domain expertise from dialogues

(Article begins on next page)

Identifying users' domain expertise from dialogues

ANONYMOUS AUTHOR(S)

Nowadays, many companies are offering chatbots and voicebots to their customers. Despite much recent success in natural language processing and dialogue research, the communication between a human and a machine is still in its infancy. In this context, dialogue personalization could be a key to bridge some of the gap, making sense of users' experiences, needs, interests and mental models when engaged in a conversation. On this line, we propose to automatically learn the user's domain expertise directly from the dialogue with the chatbot, in order to adapt its response (e.g. the complexity of the explanations) and thus improve the interaction with the user. In this paper, assuming that expertise affects linguistic features of the language, we propose a vocabulary-centered model joint with a Deep Learning method for the automatic classification of the users expertise at word- and message level. An experimentation over 5000 real conversations taken from a telco commercial chatbot carried to high accuracy scores, demonstrating the feasibility of the proposed task and paving the way for new research challenges and user-aware applications.

CCS Concepts: • **Computing methodologies → Discourse, dialogue and pragmatics**; • **Human-centered computing**;

Additional Key Words and Phrases: dialogue, deep learning, user modeling, user expertise

ACM Reference Format:

Anonymous Author(s). 1997. Identifying users' domain expertise from dialogues. In *Proceedings of ACM IUI conference (IUI'21),* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Nowadays, many companies are offering chatbot and voicebot to their customers extending omnichannel contact centers. However, despite much recent success in natural language processing and dialogue research, the communication between a human and a machine is still in its infancy [21]. Awareness of information about users can be crucial both in goal-oriented dialogues [6, 9] and chitchat settings [21] to improve the interaction. If such systems could consider users' features when engaging them in a conversation, it would improve the user satisfaction [1]. Such personal information may range from users' intents and goals [6] to users' profiles, with attributes such as home location, gender, age, profession, etc. [13], as well as users' preferences [17] and personality traits [11, 15].

On this line, we aim at making an existing CHATBOT[1] to show an intelligent behavior towards the specific user, adapting the conversation to her features, leading in such way to a more effective user experience. In particular, in this paper we focus on user's expertise in the domain and how to learn it from dialogue with the chatbot. Within the field of psychology, the definition of *expertise* has encompassed a range of ideas [4, 10], such as the "extent and organization of knowledge and special reasoning processes to development and intelligence" [8]. In general, expertise is always related to "knowledge, skill, and other cognitive concepts" [4]. In the context of this work, we consider the expertise as "the general knowledge in the telecommunications domain both at the technical level (e.g., fiber, adsl, router) and at the commercial one (e.g., commercial offers, management)".Assuming that the expertise affects the

[1]Anonymized commercial chatbot.

linguistic features of the language [19, 20], we focus on the vocabulary used by the users. State-of-the-art results [16] indicate that high-proficiency writers use words that occur less frequently in language. Accordingly, our idea is to automatically classify the user messages associating the corresponding level of expertise to the words of the message and then aggregating this information at the conversation level. The problem of assessing users' knowledge has been largely investigated in the learning domain [2], where methods for automated essay scoring by statistical analysis of linguistic features extracted from text have been proposed. Most of the works find markers of expertise by analysing linguistic features such as bag of words, concepts, negation, syntactic complexity, etc in written essays, as in [16, 20] or in chats [7, 14, 19].

In this work, we instead have a different aim. In particular, to the best of our knowledge, this is the first attempt towards the automatic identification of users expertise in dialogues with the goal of improving and enhancing the interaction with the chatbot. We tested the model with 5000 real conversations from COMPANY, reaching good accuracy levels, thus demonstrating i) the feasibility of the automatic detection of domain expertise from short messages in dialogues, and ii) the possibility to exploit such information to personalize the user interaction.

2 THE METHOD

We have turned the problem of expertise detection into a short-text classification task, focusing on the terms (lexical features) used by users. To this aim we have chosen to use neural models to analyze and predict specific classes of expertise. The advantages are many: first of all, neural networks have the great ability to generalize even over unseen data; secondly, after training, the classification of the expertise is extremely fast if compared with techniques such as cosine similarity and pattern matching. These features are essential for putting the software into production and for the real-time processing of incoming data. Furthermore, in the specific field of short-text classification, neural networks have been shown to represent the current state of the art among the existing approaches [5]. In detail, our task can be led back to the sequence labeling type (e.g, Named Entity Recognition, Semantic Role Labeling, etc.). In our case, however, labels represent levels of expertise. To manage the problem of unknown words (i.e., words that are not contained in the training set) we rely on the FastText embeddings representation [3] and the generalization power of neural networks.

Since the existing telco-oriented vocabularies only specify general and context-independent words, we manually constructed an ad hoc sentence-based annotation of 5715 terms over the 4 levels of expertise defined in Table 1. Contrary to standard vocabularies, our terms can be associated with different levels of expertise depending on the contextual sentence.

3 MODEL

3.1 Architecture

By considering the nature of the problem (sequence labeling) and the solutions already present in the literature, we chose to experiment with a bidirectional recurrent network (used as a baseline) and a bidirectional recurring network with CRF layer (current state of the art in the NER task). In our case the recurrent network architectures are GRU and LSTM. In both cases, the goal is to model the conditional probability $P(Y_1, \ldots, Y_n | X_1, \ldots, X_n)$ where Y is the desired tag sequence and X the words input sequence represented as embeddings vectors.

In the simplest case, we model the problem via a softmax function applied to the output of the neural network. In the neural layer, a forward recurrent network computes the representation of the sequence from left to right, meanwhile

Table 1. Definition and examples of expertise levels

Level 0	corresponds to the minimum set of words enabling the communication with the chatbot	internet, connection, money, top-up
Level 1	contains terms that suggest a minimum of technical knowledge on the domain. It differs from *level 0* where terms are very generic and not attributable to specific equipment or services	credit, invoice, wifi, smartphone
Level 2	technical terms appear, but still involving words in advertisements and names of commercial offers	ftth, megabit, IBAN, BIC
Level 3	corresponds to the terms associated with technical knowledge, difficult to find in posters or promotional messages including administrative terms in contractual documentation, references to regulations and resolutions	VPN, VoiP, port forwarding, penalty clause.

another backward RNN computes the same sequence in reverse. Therefore, the likelihood of every words is given by:

$$P(y|x;\theta) = \frac{exp(w[\overrightarrow{LSTM}, \overleftarrow{LSTM}] + b)}{\sum_{y' \in Y} exp(w[\overrightarrow{LSTM}, \overleftarrow{LSTM}] + b)}$$

where θ represents the LSTM parameters, including weights (w) and bias (b).

The biLSTM-CRF model, first introduced by [12], exploits the well known predictive abilities of Conditional Random Fields (CRF) networks to improve the classification of tags. Instead of considering each output tags independently, we add a CRF layer to decode the best sequence of tags, using the features given by the underlying neural network. For this purpose, we define a transition matrix $A_{i,j}$ containing the probabilities of transition from one state (tag) to the following. Therefore, we pass the output of the neural network directly to the CRF model, in this way we avoid the introduction of independence between words. Consequently, the final distribution can be represented by:

$$P(y|x;\widetilde{\theta}) = \frac{exp(\sum_i^n A_{y_{i-1},y_i}[\overrightarrow{LSTM}, \overleftarrow{LSTM}] + b)}{\sum_{y' \in Y} exp(\sum_i^n A_{y_{i-1},y_i}[\overrightarrow{LSTM}, \overleftarrow{LSTM}] + b)}$$

where $\widetilde{\theta} = \theta \cup A_{i,j}$.

For training, we minimize the negative loglikelihood $L = -\sum_i logp(y_i|X_i)$ and, for testing or decoding, we search for the optimal sequence of tags (y*) that maximizes the likelihood, or more formally $y* = arg\max_{\{y\}} p(y|X)$. This latter step can be easily accomplished by using a dynamic programming approach, such as the Viterbi's algorithm used in our work.

In this way, the biLSTM-CRF architecture is able to model both the input features (that is the semantic representation of the text) and the corresponding best tag sequence. Since there is bidirectionality in the LSTM layer, the CRF model is also able to exploit this knowledge (past and future tags) improving performance.

3.2 Expertise computation

Once we obtained the expertise for the single word that appears in the message, we aggregate those preliminary results at message and conversation level. In the computation of the final expertise score, we exploit the occurrences of the individual levels to determine a degree of confidence, in order to evaluate the goodness of the overall level. In doing so, we assume that given a sequence of various different tags, the overall expertise is the maximum value of expertise in the set of all tags. One can think, for example, of an expert user who has to describe a problem, in addition to terms that characterize his knowledge, he will also use lower level words to contextualize; however, the real expertise associated with the user is the highest level mentioned. Furthermore, a low-level user is not likely to use high-level terms by chance. Consequently this hypothesis is applied both to messages and to the entire conversation.

3.3 Confidence score

The confidence score, to be associated with expertise, is based on the entropy of the classifier output, with the aim of building a measure of "*purity*". For each message, the entropy is defined as: $H(\omega) = -\sum_{c \in C} P(\omega_c) log_2 P(\omega_c)$, where ω represents the message, c the class and $P(\omega_c)$ the probability of obtaining the tag c; this probability is calculated as $P(\omega_c) = \frac{|\omega_c|}{n_c}$, (i.e. the number of occurrences of the tag divided by the length of the message). Since the logarithm is not defined in zero, we replace this value (where present) with zero. This is justified by the limit $\lim_{n \to 0} plog(p) = 0$. It should also be noted that, since most of the tags are outsiders (i.e. they do not belong to any level), the computation must not consider this category, in this way we avoid noisy high-entropy values.

Starting from entropy (i.e. confusion or impurity), we define the confidence measure as $1 - H(\omega)$ (i.e. purity of the class). Finally, we calculate the confidence value for the entire conversation with the formula $H(\Omega) = \sum_{\omega \in \Omega} H(\omega) \frac{N_\omega}{N}$, which, in our case, simply becomes the average of the confidence values: $score(\Omega) = \sum_{\omega \in \Omega} \frac{1 - H(\omega)}{N}$, where N_ω represents the number of occurrences of the message ω and N the number of messages in the conversation. We also normalize the values thus obtained by bringing them to an interval $[0, 1]$, where 1 represents the maximum certainty on the outcome. This score is useful, in a production environment, to trigger different processes of revision and integration, especially in the cases of low values.

4 EXPERIMENTS AND RESULTS

We tested our methodology on CHATBOT's conversations (see Sec. 4.1). CHATBOT is a conversational virtual assistant offered by COMPANY, the major Telco company in COUNTRY. It is a goal-oriented dialogues chatbot devoted to support Business and Consumer customers in self-caring. CHATBOT has been trained by subject-matter experts of the Line of Business to recognize natural language requests and to dialogue about technical and commercial issues. It is available via web portal and mobile app.

4.1 Dataset

The manually annotated dataset is made up of 5000 messages coming, in equal measure, from the technical and commercial domain. The messages comprise 38290 words[2], whose distribution is the following: 4.72% level 0, 6.95% level 1, 3.15% level 2, 0.1% level3 and 85.07% outsiders. In order to train the model, the dataset is subdivided into training set (70%), validation set (15%) and test set (15%). Considering the excessive imbalance of the dataset and the problems that may arise from the under-representation of level 3, we decided to modify the distribution of classes by increasing the weight of the minority class. This has led to improvements in the accuracy of all the levels involved, since, as described in more detail in Figure 1, the classes are not independent of each other.

Referring to the initial 5000 hand-annotated messages, 1663 (33%) do not contain any annotations, meanwhile 2382 (48%) can be classified with certainty as they contain only terms belonging to the same level. It follows an average confidence score of 85.69 with only 3% messages classifiable with low confidence (e.g. score lower than 30 points). Moreover, in the 83% of cases, the choice of the label as described in Section 3.2 (via maximum) is equivalent to selecting

[2]by words we mean, in reality, even compound words; the individual tokens are 40135

the majority class (or most represented class). This means that a user categorized in a certain level will use, in most cases, words belonging to that level. Finally, we present in Figure 1 a more detailed analysis of the frequent itemsets of the label levels within messages. Two levels belong to the same itemset if they co-occur together. By observing the distribution obtained, a stronger co-occurrence of terms belonging to adjacent levels is noticeable. In particular, in a message of level *N* there is a higher probability of having terms of level *N-1* rather than terms of level *N+1*. This introduces a sort of monotonicity in the relation and confirms the hypotheses underlying the calculus in Section 3.2, according to which we can identify the user's expertise by means of the maximum-level term being used. For example one can consider level 2: it mostly appears together with the adjacent lower level 1 (241 cases), then only 102 with lower level 0, and finally it does not appear with level 3.

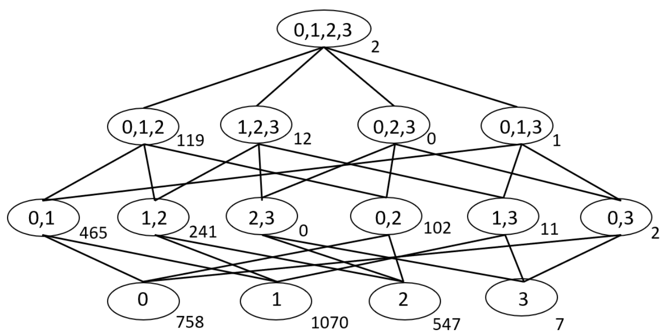

Fig. 1. Latex of the itemsets with the corresponding number of occurrences.

4.2 Classification

We have tested biLSTM, biGRU and biLSTM-CRF, biGRU-CRF models on the dataset described above. The experiments show a clear superiority of the model with Conditional Random Fields, in particular, as shown in Table 2, the architecture with GRU cells slightly outperforms the counterpart with LSTM units. With the biGRU-CRF architecture and the oversampling of the minority class (level 3), we obtain an average F1-score of 84.23. A more detailed analysis is shown in Table 3. Considering the examples of the model responses we show below, we can appreciate the differences between the first two cases, both of which can be associated with level 2, but with different scores, due to the greater variability in the second message. On the contrary, the third and fourth examples can be classified with high confidence, as there are only keywords belonging to the same level. Finally, in the latter case the entropy is maximum and it is difficult to classify the user expertise. In this situation it is possible to disambiguate the choice by analyzing the other messages in the conversation.

[Example 1] $\boxed{\textit{(Level 2, confidence 0.64)}}$ I would like to request the $\underset{level1}{\underline{\text{termination}}}$ of the $\underset{level2}{\underline{\text{TEL-YOUNG}}}$ $\underset{level1}{\underline{\text{offer}}}$ as I am finalizing the $\underset{level1}{\underline{\text{subscription}}}$ to a new $\underset{level1}{\underline{\text{plan}}}$

Table 2. F1-score of the compared models, with oversampling where indicated (overs.).

Model	F1
biLSTM	73.92
biGRU	76.22
biLSTM + CRF	78.16
biGRU + CRF	77.68
biLSTM + CRF + overs.	83.97
biGRU + CRF + overs.	**84.23**

Table 3. F1-score, Precision and Recall over expertise levels.

	Lev 0	Lev 1	Lev 2	Lev 3	outside	avg
F1	75.42	74.07	75.16	99.35	97.14	84.23
P	82.26	73.16	81.04	99.57	96.36	86.48
R	69.62	75.00	70.07	99.14	97.2	82.35

[Example 2] $\boxed{\textit{(Level 2, confidence 0.27)}}$ I want to $\underset{level0}{\underline{\text{remove}}}$ $\underset{level2}{\underline{\text{TEL-YOUNG}}}$ immediately, the other day I wasn't at home and I couldn't go to $\underset{level1}{\underline{\text{facebook}}}$ or $\underset{level1}{\underline{\text{whatsapp}}}$, so I pay for the $\underset{level0}{\underline{\text{internet}}}$ but the $\underset{level1}{\underline{\text{connection}}}$ is always slow, I want my $\underset{level0}{\underline{\text{money}}}$ back

[Example 3] $\boxed{\textit{(Level 0, confidence 1.0)}}$ home $\underset{level0}{\underline{\text{internet}}}$ is broken

[Example 4] $\boxed{\textit{(Level 3, confidence 1.0)}}$ configure $\underset{level3}{\underline{\text{port forwarding}}}$ and $\underset{level3}{\underline{\text{VPN}}}$ at the same time

[Example 5] $\boxed{\textit{(Level 2, confidence 0.5)}}$ Can I switch to $\underset{level1}{\underline{\text{fiber}}}$? I think that my area is covered by $\underset{level2}{\underline{\text{ftth}}}$ technology, can you confirm?

By looking closely at the data, we discovered the presence of some recurring patterns, for example the term "*offer*" (level 1) preceded by the proper name of the offer (level 2). These patterns can be successfully learned by the CRF model and allow the classifier to capture new commercial offers (not present during the training) and, at the same time, discard uninteresting words.

Finally, by analyzing the output of the classifier in detail, it is possible to measure the ability of the model to generalize over unseen new data. In particular we found that 34% of the new words identified (i.e. not present at the time of training) actually carry useful information (that can be associated with a level) and of these, 75% are annotated according to the correct level. Most of the discovered terms are commercial offer names (28%) or typos (6%). The characteristic of tracing new commercial offers (also belonging to other telephone operators) is essential for the correct functioning of the classifier and to keep up with the continuous evolution of the market.

5 CONCLUSION

In this paper we presented the preliminary results of the novel idea of automatic learning the users' domain expertise directly from their dialogue with a chatbot. In particular, we started from a simple model of expertise and a manual annotation of 5,000 real conversations, putting forward a neural-based classification module achieving promising results, demonstrating the feasibility of the proposal. Among all future steps, we first plan *i)* to consider other markers such as the presence of lexical and syntactical errors, anthropomorphization of the chatbot and deictic references; *ii)* to replicate the experiment with a larger training set; *iii)* to consider more complex neural mechanisms (e.g., attention [18], and *iv)* to integrate our model in CHATBOT and test the improvement of the interaction with real users.

REFERENCES

[1] Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 2 (2005), 293–327.

[2] Paulo Blikstein. 2011. Using learning analytics to assess students' behavior in open-ended programming tasks. In *Proceedings of the 1st international conference on learning analytics and knowledge*. 110–116.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[4] Lyle E Bourne Jr, James A Kole, and Alice F Healy. 2014. Expertise: defined, described, explained. *Frontiers in psychology* 5 (2014), 186.

[5] Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. 2019. Deep short text classification with knowledge powered attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6252–6259.

[6] Yun-Nung Chen, Dilek Hakkani-Tür, Gökhan Tür, Jianfeng Gao, and Li Deng. 2016. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding.. In *Interspeech*. 3245–3249.

[7] Mihai Dascalu, Erol-Valeriu Chioasca, and Stefan Trausan-Matu. 2008. ASAP-An Advanced System for Assessing Chat Participants. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. Springer, 58–68.

[8] Paul J Feltovich and Robert R Hoffman. 1997. *Expertise in context*. AAAI Press Menlo Park, CA.

[9] Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 710–718.

[10] SK Garrett, Barrett S Caldwell, EC Harris, and MC Gonzalez. 2009. Six dimensions of expertise: a more comprehensive definition of cognitive expertise for team coordination. *Theoretical Issues in Ergonomics Science* 10, 2 (2009), 93–105.

[11] Yasmín Hernández, Carlos Acevedo Peña, and Alicia Martínez. 2018. Model for personality detection based on text analysis. In *Mexican International Conference on Artificial Intelligence*. Springer, 207–217.

[12] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR* abs/1508.01991 (2015). arXiv:1508.01991 http://arxiv.org/abs/1508.01991

[13] Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503* (2017).

[14] Tayfun Kucukyilmaz, B Barla Cambazoglu, Cevdet Aykanat, and Fazli Can. 2008. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management* 44, 4 (2008), 1448–1466.

[15] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research* 30 (2007), 457–500.

[16] Danielle S McNamara, Scott A Crossley, and Philip M McCarthy. 2010. Linguistic features of writing quality. *Written communication* 27, 1 (2010), 57–86.

[17] Kaixiang Mo, Shuangyin Li, Yu Zhang, Jiajun Li, and Qiang Yang. 2016. Personalizing a dialogue system with transfer reinforcement learning. *arXiv preprint arXiv:1610.02891* (2016).

[18] Yashen Wang, He-Yan Huang, Chong Feng, Qiang Zhou, Jiahui Gu, and Xiong Gao. 2016. Cse: Conceptual sentence embeddings based on attention model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 505–515.

[19] Yonas Woldemariam, Henrik Björklund, and Suna Bensch. 2017. Predicting User Competence from Linguistic Data. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*. 476–484.

[20] Marcelo Worsley and Paulo Blikstein. 2011. What's an Expert? Using Learning Analytics to Identify Emergent Markers of Expertise through Automated Speech, Sentiment and Sketch Analysis.. In *EDM*. 235–240.

[21] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243* (2018).