ATTITUDE CLASSIFICATION IN ADJACENCY PAIRS OF A HUMAN-AGENT INTERACTION WITH HIDDEN CONDITIONAL RANDOM FIELDS

Valentin Barriere[1], *Chloé Clavel*[1], *Slim Essid*[1]

[1]LTCI, Télécom ParisTech, Université Paris Saclay, F-75013, Paris, France

ABSTRACT

In this paper, the main goal is to classify, in a human-agent interaction, the attitude of the user using hidden conditional random fields. This model allows us to capture the dynamics of the interaction in the pairs of speech turns (adjacency pairs) analyzed by our system. High level linguistic features are computed at word level. The features include syntactic features, a statistical word embedding model and subjectivity lexicons. The proposed system is evaluated on the SEMAINE corpus. We obtain a F1-score of 0.80, labeling using the most probable sequence of hidden states.

Index Terms— Hidden Conditional Random Field, Opinion Mining, Linguistic Patterns, Attitude Detection

1. INTRODUCTION

The topic of opinion mining in text has developed considerably in the last several years. The approaches to resolve those tasks are numerous, also the problem can be tackled at different levels of granularity: from fine-grained sentiment analysis using linguistic rules [1] to data-driven supervised methods [2] requiring a large amount of training labeled data. Taking the best of both worlds, hybrid approaches [3] combine the robustness and the high accuracy of Machine Learning (ML) algorithms with the stability of lexicons and linguistic patterns. In this paper, we follow the line of those studies by combining a distributed word embedding representation with lexicons, linguistic patterns and syntactic features in order to train a data-driven supervised method.

Today there are a variety of linguistic studies that carry out a very sophisticated analysis of the language (such as the Appraisal Theory [4] for the English) and some articles present adaptations of such theories that can be used in computer science [5–7]. Martin and White's [4] Appraisal Theory's framework defines the notion of *Attitude* as emotional reactions and evaluations of behaviors or things. Following [4] and the opinion representation defined in [8] for human-agent interaction, we focus on the detection and classification of attitudinal expressions in a sentence, a notion that entails affects, judgments and appreciations in the Appraisal framework [4].

The sentiment-related analysis modules integrated in Conversational Agents are general [9], not necessarily learned from conversational data and generally not shaped for a conversational context. Nevertheless, some work has been done in that special context. Langlet and Clavel [7] combine rules with complex linguistic patterns distributed over the different speakers to detect the likes and dislikes of the user in a human-agent interaction. They integrate a basic level of dialogic context: the adjacency pairs (APs) which are **a pair composed of two utterances by two speakers, one after the other**. We can also count ML based methods [10] using the agent's dialog acts as input features of the classifier to detect the emotional state of a child during a human-agent interaction.

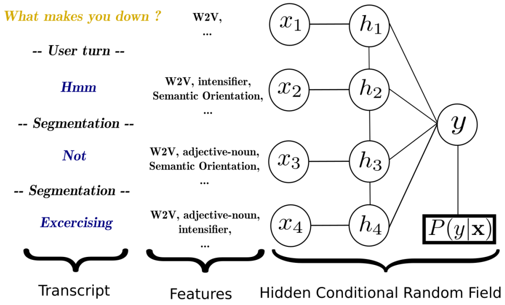
Thus, we investigate a series of statistical graphical latent state models in order to model the user's attitude in the interactional context of an AP inside a human-agent interaction. Based on the idea that speech is sequential, and that an attitude can be expressed over the AP and can be modeled by a sequence of hidden states, the user's expression of attitude depending on the agent's utterance, we use Hidden Conditional Random Fields (HCRF). It is a variant of the well-known Conditional Random Fields (CRF) which have the advantages to be discriminative and interpretable. Besides, they do not require a lot of data during the training phase. The HCRF model has been successfully used to analyze sequences of text, audio or visual data to be labeled globally with only one output [11]. Latent state models have already proven their efficiency for sentiment analysis [12, 13] or agreement classification [14].

We previously showed [13] that HCRF was able to model the intra-speaker dynamics of opinion in transcripts from a non-interactional context. In this work, we consider the task of labeling an audio transcript of a dyadic interaction between an agent and a user with respect to the user's attitudes. We firstly chose to segment the pair of speech turns at the word level in order to fully use the syntactic relations between close words. Then we try different configurations in order to take into account the interactional context of the human-agent dyad. The objective here is to investigate the potential of HCRF for a classification using transcripts from oral interactions. The discriminative nature of CRF will enable some strong linguistic rules combined with other features to emerge directly from the learning phase.

In the second section of this paper, we will present our classifier, the features we chose for our task and the different

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

Fig. 1. Overview of the system, zoom on Model 3



interactional models. In the third section, we will present the dataset, talk about our experiments and results and finish in the fourth section with a discussion of the results and then we will conclude our paper.

2. FEATURE AND CLASSIFICATION MODELS DESCRIPTION

2.1. General Framework

The main idea is firstly to segment the text into relevant units (we will see further in Figure 2, depending on the interaction models). Secondly, to extract the features from each unit (Subsection 2.2) and finally use them to feed the HCRF (Subsection 2.3). After the training phase, the model can predict the most probable label for an unseen AP. Different types of text segmentation and training for the HCRF weights, corresponding to different representations of the dyad have been designed. The different configurations are presented in Figure 1. In Subsection 2.4, we present the different HCRF interaction models we investigated in this paper, meaning the different ways to integrate the descriptors into features with respect to the segmentation and the way we trained the weights of the HCRF. An overview of the different architectures is available in Figure 2, with a zoom on the 3rd interaction model (HCRF-3) in Figure 1.

2.2. Features

We used a set of features that was designed and tested in a previous work for a sentiment analysis task [13] without considering the interactional context. In this work, the same feature set was used for both the agent and the user.

We can sort the textual features we use into 3 groups :
- *The N-grams* with the Bag-of-N-grams (BoNG);
- *The distributed representations* with word2vec [15];
- *The linguistic and lexicon-based features* with the linguistic patterns using subjectivity lexicons and the syntactic features.

When we had to integrate the descriptors on an utterance, we averaged the word-vectors before normalizing on each dimension and we used the number of linguistic patterns we detected as well as the scores from the subjectivity lexicons for every word. We standardized each linguistic feature. Due to the space limit, all the further details on the feature set can be found in [13].

2.3. Classification Model

The HCRF [11] allows to map a sequence of observation $\mathbf{x} = \{x_1, ..., x_L\}$ to a label y, using the sequences of hidden variables $\mathbf{h} = \{h_1, ..., h_L\}$. For that it uses the compatibility between the observations x_j and the hidden state h_j, the compatibility between the hidden state and the global label, and the compatibility to goes from one hidden state to another: terms 1, 2 and 3 from Equation (1) with weights θ_o, θ_s and θ_t associated. The weights are the parameters of the model to optimize.

$$\Psi(y, \mathbf{h}, \mathbf{x}; \theta) = \sum_j \langle \phi(x_j) \,|\, \theta_o(h_j) \rangle \\ + \sum_j \theta_s(y, h_j) + \sum_j \theta_t(y, h_j, h_{j+1}) \quad (1)$$

Then, the decision is usually made using the the posterior probability $P(y|\mathbf{x}, \theta)$ (Equation (2)). We also investigated to use the label of the most probable hidden sequence $y* = \arg\max_y \max_{\mathbf{h}} P(y, \mathbf{h}|\mathbf{x}, \theta)$.

$$P(y|\mathbf{x}, \theta) = \sum_{\mathbf{h}} P(y, \mathbf{h}|\mathbf{x}, \theta) = \frac{\sum_{\mathbf{h}} e^{\Psi(y,\mathbf{h},\mathbf{x};\theta)}}{\sum_{y',\mathbf{h}} e^{\Psi(y',\mathbf{h},\mathbf{x};\theta)}}, \quad (2)$$

The model is classically trained by minimizing an ℓ_2-norm regularized negative log-likelihood cost [11].

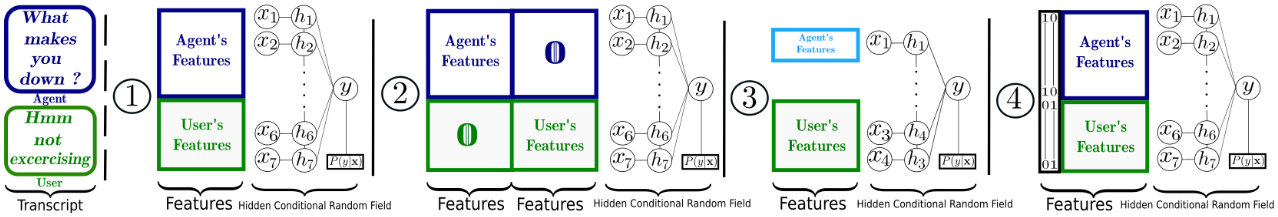
2.4. Presentation of the Different Interaction Systems

We tried different segmentations of the text and integrations of the features in the model in order to take advantage of the interactional context. Detecting the user's attitudes during a human-agent conversation does not need to pay as much attention to the agent utterance as it does to the user's one. As a baseline, we took only into account the user's utterance (model **HCRF-0**). Then we studied four different models in order to model the interactional context of the AP. Firstly, we use a simple way to integrate the features into the HCRF model: segmenting the sentence word by word and training and sharing the same weights θ_o between the user and the agent (see Figure 2, model **HCRF-1**). In the model **HCRF-2**, following the idea that the roles of the agent and user were not symmetric in this task, we trained different weights θ_o for their respective features. In the model **HCRF-3**, we decided to integrate all the features of the agent over its whole utterance to get one vector, we wanted the agent to have less impact on the final decision. However, we wanted it to influence the beginning of the latent variable sequence. Finally, for the model **HCRF-4** we used a new feature indicating if the agent or the user was speaking, as a fusion of the models 1 and 2.

3. EXPERIMENTS AND RESULTS

We tested 5 models with different feature sets and segmentations in order to validate our interaction models. Firstly,

Fig. 2. The different HCRF configurations investigated to model the human-agent interaction



we created a baseline for our task using a logistic regression model with Bag-of-N-Gram features following the protocol of [16] for a binary sentiment analysis task. Then we tried with our feature set (see Section 2.2). In order to take into account the asymmetry of the interaction, we tried 4 different feature configurations (see 3.2). Logistic regression does not take into account the dynamics of the observations, which is important for speech data. We then used a sequential model as second baseline, in order to compare with a state-of-the-art model system : a recurrent neural network (RNN-LSTM) [17]. Compared to these models, HCRF allows for improved interpretability while requiring less data for the training and model the dynamics of opinion-related phenomena (emotional states, stances, etc.) through latent states.

In order to validate our model, we used a 10-fold Cross-Validation (CV) where train and test sets are disjoint and do not contain any data from a common session (valmajnoting that it can contain the same speakers since there is very few operators).

3.1. Dataset

In this study, we used two different subsets of APs extracted from the SEMAINE corpus [18] which is a corpus of dyadic interactions between a human user and a human operator playing the role of a virtual agent that has a well defined attitude: Poppy, happy and outgoing, Obadiah, depressed and gloomy, Spike, angry, and Prudence shy and sensible. The verbal content has been annotated with respect to the user's attitudes following the scheme described in [7, 19] which has been built from the appraisal theory of Martin and White [4]. In this theory, Attitude refers to the emotional reactions and the evaluations of behaviors or things. The valence of each attitude expression is also given (positive or negative). We used the APs annotated by Langlet and Clavel [7] using the Amazon Mechanical Turk and used for the evaluation of their system. In order to augment the dataset, we extended the set with APs from SEMAINE annotated by a linguist in attitude with associated valence, following the same scheme.

We chose to discard the APs containing at least 2 user's attitudes with opposite valences in order to have a real binary classification task. This led us to a total of 958 APs (145 negatives, 534 neutrals, 279 positives). For a binary classification task, this corresponds to 424 APs and **8880** words. The attitude annotations were made without using the audiovisual information, which is consistent with the system provided in this paper which relies only on text features.

3.2. Baselines using LogReg and LSTM

Methodology : We considered several baseline models with a simple textual feature set and our feature set that we tested for different levels of textual representation (at the AP level or using each word as an observation). We firstly used a Logistic Regression model as a simple baseline with different strategies to take into account the interaction between the agent and the user. For the Logistic Regression, we used 4 configurations. We integrated the features over the whole AP or over the agent utterance or speaker utterance, or over both, and concatenated the result. We then changed for a more sophisticated feature set, that is a representation using the statistical word embedding model from [20] described in 2.2. After a tokenization[1] we used a spell checker[2] to eliminate the numerous typos from the transcription and to clean the text before taking the word-vectors. We tried different strategies to integrate the word embeddings on the AP, using 4 functionals (the average, the median, the maximum and the minimum) to obtain one vector of the same size, and normalized them. In order to help the determination of the opinion we added the *linguistic and lexicon-based feature set* (as described in 2.2).For the LSTM, we concatenated the features of the agent and the user segmenting word by word, like for the HCRF-1 (Figure 2). Regarding the tuning of the hyperparameters and the implementation of the models, we used the same experimental methodology as in our previous work [13] and made every possible efforts to train the LSTM model. We envisaged doing fine-tunning but data was lacking since the task we consider is atypical.

Results : The results of the baselines are listed in the first part of Table 1 using F1-scores and accuracy. In this table, the global $F1$ (the harmonic mean of recall and precision) is the average $F1$ of both classes ($F1+$ and $F1-$) and $Accuracy$ is the percentage of true predictions. As expected the BoNG representation is not suited for short documents such as the user's utterance or AP [22].The BoNG model struggles to be

[1] We used the CoreNLP from Standford [21]

[2] https://github.com/phatpiglet/autocorrect

Table 1. F1-scores and accuracies results with different feature sets and models

Features	Model	F1+	F1-	F1	Acc
Majority label	Dummy	79	0	40	66
BoNG	LogReg	73	36	55	62
Our set	LogReg	70	39	55	60
Our set	LSTM	83	64	74	77
BoNG	HCRF-1	80	55	67	72
Our set	HCRF-0	82	67	74	77
Our set	HCRF-1	84	74	79	80
Our set	HCRF-2	84	73	78	80
Our set	HCRF-3	**86**	**73**	**79**	**81**
Our set	HCRF-4	**86**	**75**	**80**	**82**

efficient at classifying negative APs. Using our set which contains sentiment-related and linguistic features brings an improvement over this. The results of the RNN-LSTM are better for the negative class. Though it has the potential to capture some dynamics, the neural network requires more data than available in the considered corpus to be fully effective.

3.3. HCRF models

Methodology : The existence of latent states in HCRF makes them useful to model a dynamic system like, for example, the emotional state of the speaker. Using our feature set, including sentiment-related features and a distributed representation, the model is expected to use the information contained inside the already-trained vectors and exploit more effectively the concepts employed by the speakers. We used the 5 configurations presented in Figure 2 and explained in the section 2.4. Regarding the tuning of the hyperparameters and the implementation of the models, we used the same experimental methodology as in previous work [13].

Results : The results with the HCRF models are summarized in Table 1. Using the global label of the most probable sequence improved the results compared to a classical approach. The best F1-score was reached using the fourth configuration with 6 hidden states, a ℓ_2 regularization coefficient equal to 0.05 and a context window of size 2. This configuration is taken into account the dynamics of interaction by integrating the speech turn of the agent into a unique observation before using the words of the user one by one (see Figure 2). Our feature set improves the F1 score on the negative files for the HCRF-1 going from 55 to 69. We compare our approach with a regular sentiment analysis model taking only into account the speaker, and can see the improvements over the model HCRF-0: the F1-score going from 74 to 80. The HCRF are particularly good on the negative files when compared to the baseline. Though a 10-fold CV over such a dataset does not allow one to conclude about the statistical significance of this difference in performance, the results are believed to be very promising and efforts are currently being made to collect and annotate more data to obtain a more solid validation.

Fig. 3. Example of a tagged AP (green/red means the state compatible with positive/negative label)

Agent: That's good
User: I don't like this weather

4. DISCUSSION

The best system is the one simply obtained using an indicator-feature expressing who is the speaker. The Viterbi labeling improve greatly the performances over the negative files, outclassing the F1-score of LSTM baseline over the negative files by 11 points (from 64 to 75). Sharing the weights between the two speakers seems to be the best option. The dataset is too tiny too learn two times more weights leading to lowest results for the model 2. Nevertheless, differentiating the two speakers by special features or a special integration of the features like for the models 3 and 4 is a plus, leading to the best results. The model 0 is not very good on the negative files, so using the context of the AP is very important. Indeed, in several negative files Spike is propagating its negative attitude to the user. The user's answers could be considered neutral without the context, like "*you do too*" or "*yeah I've getting*".

When we analyzed the weights of the system, we remarked that 2 of the hidden states were really compatible with the positive or with the negative labels with highly positive θ_s, the other were neutrals. The weights associated with those states were also very specific. The observation weights θ_o with the biggest values for the state highly compatible with positive (respectively negative) label were positive values (respectively negative) from subjectivity lexicons. For the neutral states, it was neutral values from the subjectivity lexicons (i.e. if a word is neutral, like *table*). When we looked at the Viterbi labeling[3] it was possible to visualize the changes in opinion of the speaker during its utterance (see Figure 3). Moreover, it is important to note that the indicator of the user was highly compatible with the positive and negative states while not with the neutral ones, explaining the improvements of HCRF-4 over HCRF-1.

5. CONCLUSION AND FUTURE WORK

In this paper, we have presented HCRF models that use the interactional context in order to detect a user's attitude in a human-agent interaction. Our textual feature set includes word embeddings, linguistic rules and clues from a subjectivity lexicon. The use of HCRF classifiers allows us to implicitly learn local linguistic representations of each word of the transcript in order to model a dynamic process. We investigated models that take benefits of the interactional context in order to analyze the user of a human-agent interaction.

In our future work, it could be useful to investigate the use of agent specific features like dialog acts and more complex distributed linguistic patterns. Further, we would like to augment the size of the dataset in order to obtain significant results.

[3] Labeling $\mathbf{h*} = \arg\max_{\mathbf{h}} P(\mathbf{h}|y, \mathbf{x}, \theta)$

6. REFERENCES

[1] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede, "Lexicon-Based Methods for Sentiment Analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[2] Richard Socher, Alex Perelygin, and Jy Wu, "Recursive deep models for semantic compositionality over a sentiment treebank," *EMNLP-2013: Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

[3] Bishan Yang and Claire Cardie, "Joint Inference for Fine-grained Opinion Extraction," *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,*, pp. 1640–1649, 2013.

[4] James R Martin and Peter R R White, "The Language of Evaluation: The Appraisal Framework," *Lecture Notes in Computer Science*, p. 256, 2003.

[5] Janyce Wiebe, Theresa Wilson, and Claire Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005.

[6] Jonathon Read and John Carroll, "Annotating expressions of Appraisal in English," *Language Resources and Evaluation*, vol. 46, no. 3, pp. 421–447, 2012.

[7] Caroline Langlet and Chlo Clavel, "Improving social relationships in face-to-face human-agent interactions: when the agent wants to know user's likes and dislikes," in *ACL-IJCNLP 2015*, 2015, pp. 1064–1073.

[8] Caroline Langlet and Chlo? Clavel, "Grounding the detection of the user's likes and dislikes on the topic structure of human-agent interactions," *Knowledge-Based Systems*, vol. 106, pp. 116–124, 2016.

[9] Chloe Clavel and Zoraida Callejas, "Sentiment Analysis: From Opinion Mining to Human-Agent Interaction," 2016.

[10] Serdar Yildirim, Shrikanth Narayanan, and Alexandros Potamianos, "Detecting emotional state of a child in a conversational computer game," *Computer Speech and Language*, vol. 25, no. 1, pp. 29–44, 2011.

[11] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell, "Hidden conditional random fields.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 10, pp. 1848–1853, 2007.

[12] Louis-philippe Morency, Rada Mihalcea, and Payal Doshi, "Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web," *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI-11)*, pp. 169–176, 2011.

[13] Valentin Barriere, Chloe Clavel, and Slim Essid, "Opinion Dynamics Modeling for Movie Review Transcripts Classification with Hidden Conditional Random Fields," in *INTERSPEECH*, 2017.

[14] Konstantinos Bousmalis, Louis-Philippe Morency, and Maja Pantic, "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, 2011, pp. 746–752.

[15] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pp. 1–12, 2013.

[16] Martin Wöllmer, Felix Weninger, Tobias Knaup, Bjorn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency, "YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.

[17] Sepp Hochreiter and J Urgen Schmidhuber, "LONG SHORT-TERM MEMORY," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] Gary McKeown, Michel F. Valstar, Roderick Cowie, and Maja Pantic, "The semaine corpus of emotionally coloured character interactions," *2010 IEEE International Conference on Multimedia and Expo, ICME 2010*, pp. 1079–1084, 2010.

[19] Caroline Langlet and Chloe Clavel, "Modelling user's attitudinal reactions to the agent utterances : focus on the verbal content," in *5th International Workshop on Corpora for Research on Emotion, Sentiment & Social Signals (ES3 2014).*, Reykjavik, Iceland, 2014.

[20] Tomas Mikolov, I. Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Proc. NIPS*, 2013, pp. 1–9.

[21] Sebastian Schuster and Christopher D. Manning, "Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks," *Proceedings of LREC 2016*, pp. 2371–2378, 2016.

[22] Farah Benamara, Maite Taboada, and Yannick Mathieu, "Evaluative Language Beyond Bags of Words: Linguistic Insights and Computational Applications," *Computational Linguistics*, pp. 201–264, 2016.