Detecting Age-Related Linguistic Patterns in Dialogue: Toward Adaptive Conversational Systems

Lennert Jansen[1], **Arabella Sinclair**[1], **Margot J. van der Goot**[2], **Raquel Fernández**[1], **Sandro Pezzelle**[1]

[1]Institute for Logic, Language and Computation (ILLC), University of Amsterdam
[2]Amsterdam School of Communication Research (ASCoR), University of Amsterdam
lennertjansen95@gmail.com
{a.j.sinclair|m.j.vandergoot|raquel.fernandez|s.pezzelle}@uva.nl

Abstract

This work explores an important dimension of variation in the language used by dialogue participants: their age. While previous work showed differences at various linguistic levels between age groups when experimenting with written *discourse* data (e.g., blog posts), previous work on *dialogue* has largely been limited to acoustic information related to voice and prosody. Detecting fine-grained linguistic properties of human dialogues is of crucial importance for developing AI-based conversational systems which are able to adapt to their human interlocutors. We therefore investigate whether, and to what extent, current text-based NLP models can detect such linguistic differences, and what the features driving their predictions are. We show that models achieve a fairly good performance on age-group prediction, though the task appears to be more challenging compared to discourse. Through in-depth analysis of the best models' errors and the most predictive cues, we show that, in dialogue, differences among age groups mostly concern stylistic and lexical choices. We believe these findings can inform future work on developing controlled generation models for adaptive conversational systems.

1 Introduction

Research on developing conversational agents has experienced impressive progress, particularly in recent years (McTear, 2020). However, artificial systems that can tune their language to that of a particular individual or group of users continue to pose more of a challenge. Recent examples of this line of research include adaptation at style level (Ficler and Goldberg, 2017), persona-specific traits (Zhang et al., 2018), or other traits such as sentiment (Dathathri et al., 2020).

age 19-29
A: oh that's cool
B: different sights and stuff
A: oh

age 50+
A: well quite and I'd have to come back as well
B: that's of course
A: and make up for you know

Figure 1: Example dialogue snippets from speakers of different age groups in the British National Corpus. We conjecture that stylistic and lexical differences between age groups can be detected. Here, we experiment at the level of the utterance.

Personalised interaction is of crucial importance to obtain systems that can be trusted by users and perceived as natural (van der Goot and Pilgrim, 2019), but most of all to be accessible to varying user profiles, rather than targeted at one particular user group (Zheng et al., 2019; Zeng et al., 2020).

In this work, we focus on one particular aspect that may influence conversational agent success: user age profile. We investigate whether the linguistic behaviour of conversational participants differs across age groups using state-of-the-art NLP models on purely textual data, without considering vocal cues. We aim to detect age from characteristics of language use and adapt to this signal, rather than work from ground-truth metadata about user demographics. This is in the interest of preserving privacy, and from the perspective that while age and language use may have a relationship, this will not be linear (Pennebaker and Stone, 2003) and there are individual differences.

Previous work on age detection in dialogue has focused on speech features, which are known to systematically vary across age groups. For example, Wolters et al. (2009) learn logistic regression age classifiers from a small dialogue dataset using different acoustic cues supplemented with a small set of hand-crafted lexical features, while Li et al. (2013) develop SVM classifiers using acoustic and prosodic features extracted from scripted utterances spoken by participants interacting with an artificial system. In contrast to this line of work, we investigate whether different age groups can be detected from textual linguistic information rather than voice-related cues. We explore whether, and to what extent, various state-of-the-art NLP models are able to capture such differences in dialogue data as a preliminary step to age-group adaptation by conversational agents.

We build on the work of Schler et al. (2006), who focus on age detection in written discourse using a corpus of blog posts. The authors learn a Multi-Class Real Winnow classifier leveraging a set of pre-determined style- and content-based features, including part-of-speech categories, function words, and the 1000 unigrams with the highest information gain in the training set. They find that content features (lexical unigrams) yield higher accuracy (74%) than style features (72%), while their best results (76.2%) are obtained with their combination. We extend this investigation in several key ways: (1) we leverage state-of-the-art NLP models that allow us to learn representations end-to-end, without the need to specify concrete features in advance; (2) we apply this approach to dialogue data, using a large-scale dataset of transcribed, spontaneous open-domain dialogues, and also use this approach to replicate the experiments of Schler et al. (2006) on disccourse; (3) we show that text-based models can indeed detect age-related differences, even in the case of very sparse signal at the level of dialogue utterances; and finally (4) we carry out an in-depth analysis of the models' predictions to gain insight on which elements of language use are most informative.[1]

Our work can be considered a first step toward the modeling of age-related linguistic adaptation by AI conversational systems. In particular, our results can inform future work on controlled text generation for dialogue agents (Dathathri et al., 2020; Madotto et al., 2020).

age	**#samples**	**#tokens**	**mean L (± sd)**	**min-max L**
19-29	33,641	381,195	11.3 (±15.98)	1-423
50+	33,641	406,157	12.1 (±21.62)	1-1246
all	67,282	787,352	11.7 (±19.0)	1-1246

Table 1: Descriptive statistics of the dataset. L means length, i.e., number of tokens in a sample.

2 Data

We use a dataset of dialogue data where information about the age of the speakers involved in the conversation is available (see the dialogue snippets in Figure 1), i.e., the spoken partition of the British National Corpus (Love et al., 2017). This partition includes spoken informal open-domain conversations between people that were collected between 2012 and 2016 via crowd-sourcing, and then recorded and transcribed by the creators. Dialogues can be between two or more interlocutors, and are annotated along several dimensions including age and gender together with geographic and social indicators. Speaker ages are categorized in ten brackets: 0-10, 11-18, 19-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, and 90-99.

We focus on conversations that took place between two interlocutors, and only consider dialogues between people of the same age group. We then restrict our investigation to a binary opposition: *younger* vs. *older* age group. We split the dialogues into their constituent utterances (e.g., from each dialogue snippet in Figure 1 we extract three utterances), and further pre-process them by removing non-alphabetical characters. Only samples which are not empty after pre-processing are kept. For the *younger* group, we consider the 19-29 bracket, which contains 138,662 utterances. For the *older*, we group conversations from five brackets: 50-59, 60-69, 70-79, 80-89, and 90-99 (hence, 50+), which sums up to a total of 33,641 utterances. The choice of grouping these brackets is a trade-off between experimenting with fairly distinct age groups (the age difference between them is at least 20 years) and obtaining large-enough data for each of them.

We randomly sample 33,614 utterances from the 19-29 group in order to experiment with a balanced number of samples per group. The resulting dataset, that we use for our experiments, includes around 67K utterances with an average length of 11.7 tokens. Descriptive statistics are in Table 1.

[1] Code and data available at: `https://github.com/lennertjansen/detecting-age-in-dialogue`

3 Method

We frame the problem as a binary classification task: given some text, we seek to predict whether the age class of its speaker is *younger* or *older*.

3.1 Models

We experiment with various models, that we briefly describe below. Details on model training and evaluation are given at the end of the section.

***n*-gram** Our simplest models are based on n-grams, which have the advantage of being highly interpretable. Each data entry (i.e., a dialogue utterance) is split into chunks of all possible contiguous sequences of n tokens. The resulting vectorized features are used by a logistic regression model to estimate the odds of a text sample belonging to a certain age group. We experiment with unigram, bigram and trigram models. A bigram model uses unigrams and bigrams, and a trigram model unigrams, bigrams, and trigrams.

LSTM and BiLSTM We use a standard Long Short-Term Memory network (LSTM) (Hochreiter and Schmidhuber, 1997) with two layers, embedding size 512, and hidden layer size 1024. Batch-wise padding is applied to variable length sequences. The original model's bidirectional extension, the bidirectional LSTM (BiLSTM) (Schuster and Paliwal, 1997), is also used. Padding is similarly applied to this model, and the following optimal architecture is experimentally found: embedding size 64, 2 layers, and hidden layer size 512. Both RNN models are found to perform optimally for a learning rate of 10^{-3}.

BERT We experiment with a Transformer-based model, i.e., BERT (Devlin et al., 2019). BERT is pre-trained to learn deeply bidirectional language representations from massive amounts of unlabeled textual data. We experiment with the base, uncased version of BERT, in two settings: by using its pre-trained frozen embeddings (BERT$_{frozen}$) and by fine-tuning the embeddings on our age classification task (BERT$_{FT}$). BERT embeddings are followed by dropout with probability 0.1 and a linear layer with input size 768.

Model	**Accuracy** ↑ better	$\boldsymbol{F}_1^{(19-29)}$ ↑ better	$\boldsymbol{F}_1^{(50+)}$ ↑ better
Random	0.500	0.500	0.500
unigram	0.701 (0.007)	0.708 (0.009)	0.693 (0.004)
bigram	0.719 (0.002)	0.724 (0.003)	0.714 (0.003)
trigram	0.722 (0.001)	0.727 (0.003)	0.717 (0.001)
LSTM	0.693 (0.003)	0.696 (0.005)	0.691 (0.007)
BiLSTM	0.691 (0.009)	0.702 (0.017)	0.679 (0.007)
BERT$_{frozen}$	0.675 (0.003)	0.677 (0.008)	0.673 (0.010)
BERT$_{FT}$	**0.729** (0.002)	**0.730** (0.011)	**0.727** (0.010)

Table 2: Test set results averaged over 5 random initializations. Format: *average metric (standard error)*. Values in **bold** are the highest in the column; in blue, the second highest.

Experimental details The dataset is randomly split into a training (75%), validation (15%), and test (10%) set. Each model with a given configuration of hyperparameters is run 5 times with different random initializations. All models are trained on an NVIDIA TitanRTX GPU.

The n-gram models are trained in a One-vs-Rest (OvR) fashion, and optimized using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm (Liu and Nocedal, 1989), with a maximum of 10^6 iterations. The n-gram models are trained until convergence or for the maximum number of iterations.

LSTMs and BERT models are optimized using Adam (Kingma and Ba, 2015), and trained for 10 epochs, with an early stopping patience of 3 epochs. The RNN-based models' embeddings are jointly trained, and optimal hyperparameters (i.e., learning rate, embedding size, hidden layer size, and number of layers) are determined using the validation set and a guided grid-search. BERT$_{FT}$ is fine-tuned on the validation set for 10 epochs, or until the early stopping criterion is met. BERT has a maximum input length of 512 tokens. Sequences exceeding this length are truncated.

4 Results

We report accuracy and F_1 for each age group in Table 2. As can be seen, the performance of all models is well beyond chance level, which indicates that age-related linguistic differences can be detected, to some extent, even by a simple model based on unigrams. At the same time, BERT fine-tuned on the task turns out to be the best-performing model both in terms of accuracy (0.729) and F_1 scores, which confirms the effectiveness of Transformer-based representations to encode fine-grained linguistic differences. However, it can be noted that the model based on tri-

	% cases	**avg. length (±std)***
both correct	63.17%	13.51 (±18.98)
both wrong	19.78%	5.82 (±8.33)
only trigram correct	7.91%	10.44 (±11.66)
only BERT correct	9.14%	11.53 (±12.12)

Table 3: Percentage cases of (non-)overlapping (in)correctly predicted cases between trigram and BERT$_{FT}$. *Utterance length measured in tokens.

grams is basically on par with BERT in terms of accuracy (0.722), and well above both the LSTM and BiLSTM models (0.693 and 0.691, respectively). A similar pattern is observed for F_1 scores, where BERT$_{FT}$ and the trigram model achieve comparable performance, with LSTMs being overall behind.

Overall, our results indicate that text-based models are effective, to some extent, in predicting the age group to which a speaker involved in a dialogue belongs. This complements previous evidence that age-related features can be detected in discourse (Schler et al., 2006), and shows that in dialogue the task appears to be somehow more challenging: The improvement in accuracy with respect to the majority/random baseline is lower in our dialogue results (+22.9%) as compared to what observed in discourse both by Schler et al. (2006) (+32.4%) and by us (+27%) when replicating their study using the models and experimental setup described in Section 3.1. Similarly to dialogue, BERT$_{FT}$ achieves the highest results in discourse (0.742). In contrast, both LSTMs (0.663) and n-grams (0.625) significantly lag behind it. Note that, although based on the same corpus of texts, i.e., the Blog Authorship Corpus,[2] and the same 3 age groups, i.e., 13-17, 23-27, and 33+, our replicated results are not fully comparable to those by Schler et al. (2006). Due to our more cautious data pre-processing, we experiment with more samples than they do (677K vs. 511K), which in turn leads to a different majority baseline.

There can be several reasons why age group detection is more challenging in dialogue than in discourse. For example, in dialogue there may be dimensions of variation, such as turn-taking patterns, that are not captured by our models and experimental setup. Yet, the present results do reveal a few interesting insights. In particular, the very good performance of the trigram model suggests that leveraging 'local' linguistic features captured by n-grams is extremely effective in *dialogue*. This could indicate that differences among various age groups are at the level of local lexical constructions. This deserves further analysis, that we carry out in the next section.

5 Analysis

We compare the two best-performing models, i.e., BERT$_{FT}$ and the one using trigrams, and aim to shed light on what cues they use to solve the task. We first compare the prediction patterns of the two models, which allows us to detect easy and hard examples. Second, we focus on the trigram model and report the n-grams that turn out to be most informative to distinguish between age groups.

5.1 Comparing Model Predictions

We split the data for analysis by whether or not both models make the same correct or incorrect prediction, or whether they differ. Table 3 shows the breakdown of these results. As can be seen, a quite large fraction of samples are correctly classified by both models (63.17%), while in 19.78% cases neither of the models make a correct prediction. The remaining cases are almost evenly split between cases where only one of the two is correct. As shown in Figure 2, the 19-29 age group appears to be be slightly easier compared to the 50+ group, where models make more errors.

To qualitatively inspect what the utterances falling into these classes look like, in Table 4 we show a few cherry-picked cases for each age group. We notice that, not surprisingly, both models have trouble with backchanneling utterances consisting of a single word, such as *yeah*, *mm*, or *really?*, which are used by both age groups. For example, both models seem to consider *yeah* as a 'young' cue, which leads to wrong predictions when *yeah* is used by a speaker in the 50+ group. As for the utterance *really?*, BERT$_{FT}$ assigns it to the 50+ group, while the trigram model makes the opposite prediction. This indicates that certain utterances simply do not contain sufficient distinguishing information, and model predictions that are based on them should therefore not be considered reliable. This seems to be particularly the case for short utterances. Indeed, through comparing the average length of the utterances incorrectly classified by both models (rightmost column

[2] The corpus contains blog posts appeared on `https://www.blogger.com`, gathered in or before August 2004.

age	both correct	both wrong	only BERT$_{FT}$ correct	only trigram correct
19-29	I don't know? sounded crazy	that's a lot of people for one house	yeah okay	really?
19-29	yeah	well there you go	oh I'm not very good at that	I've got a pen I've got a pen
19-29	do you have exams again?	mm	empty promises isn't it?	day of death and ice-cream
50+	and as I say	yeah	really?	well if I were you
50+	yes	that would be controversial	yeah it seems to	that's it
50+	oh really?	he's got that already	that we caused it	oh I thought you said Godzilla

Table 4: Examples where both models are correct/wrong or only BERT$_{FT}$/trigram is correct.

of Table 3), we notice that they are much shorter than those belonging to the other cases. This is interesting, and indicates a key challenge in the analysis of dialogue data: on average, shorter utterances contain less signal. On the other hand, short utterances can provide rich conversational signal in dialogue; for example, backchanneling, exclamations, or other acknowledging acts. As a consequence, using length alone as a filter is not an appropriate approach, as it can remove aspects of language use key to differentiating speaker groups.

5.2 Most Informative N-grams

Analyzing the most informative n-grams used by the trigram model allows us to qualitatively compare the linguistic differences inherent to each age group. In Table 5 we report the top 15 n-grams per group. We find, firstly and intuitively, that colloquial language seems somewhat generational, with unigrams particularly indicative of younger speakers consisting of words such as *cool* and *massive*, and for older speakers, words like *wonderful*. These unigrams are both informative to the model and indicative of differences in both formality and 'slang' use across age groups.

These most informative n-grams also indicate differences in back-channeling use between age groups; younger speaker's language is more characterized by the use of *um*, *hmm*, while the top n-grams in the older category will more likely use *yes*, *right*, *right right*. A feature of younger language also apparent from these examples is in their use of more informal language, which also extends to the use of foul language, making up a percent of the most informative unigrams shown in Table 5.

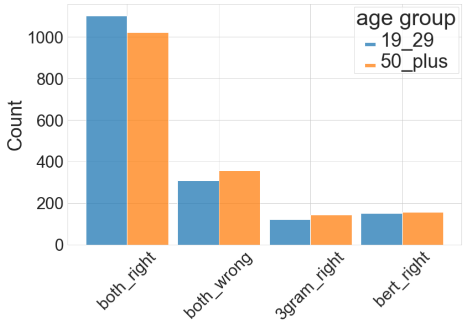

Figure 2: Distribution of predicted cases by trigram and BERT$_{FT}$ models, split by age groups.

Interestingly, while topic words make up many of the most informative n-grams for older speakers in Table 5, younger speakers are more defined by their use of slang words such as *wanna*, foul language, or adjectives such as *cute*, *cool*, and *massive*. A key finding from Schler et al. (2006) is in the sentiment of language playing an important role, something which some of the most informative n-grams suggest may also be true for the dialogue dataset. As Table 5 demonstrates, younger speakers use more dramatic language such as negative foul words, and positive *love, cute, cool*; all words with a strong connotative meaning. We believe that further inspection is needed to determine whether the same sentiment pattern will be true of

	19-29		50+
coef.	**n-gram**	**coef.**	**n-gram**
-3.20	um	2.37	yes
-2.84	cool	2.12	you know
-2.58	s**t	2.09	wonderful
-2.12	hmm	1.90	how weird
-2.09	like	1.84	chinese
-2.02	was like	1.73	right
-1.96	love	1.71	building
-1.96	as well	1.66	right right
-1.88	as in	1.55	so erm
-1.84	cute	1.43	mm mm
-1.82	uni	1.41	cheers
-1.79	massive	1.39	shed
-1.79	wanna	1.37	pain
-1.79	f**k	1.36	we know
-1.72	tut	1.08	yeah exactly

Table 5: Top 15 most informative n-grams per age group used by the trigram model. **coef.** is the coefficient (and sign) of the corresponding n-gram for the logistic regression model: the higher its absolute value, the higher the utterance's odds to belong to one age group. * indicates foul language.

dialogue as it has been reported to be in discourse.

6 Conclusion

We investigated whether, and to what extent, NLP models can detect age-related linguistic features in dialogue data. We showed that, in line with what we observed for discourse, state-of-the-art models are capable of doing so with a reasonable accuracy, in particular when the dialogue fragment is long enough to contain discriminative signal. At the same time, we found that much simpler models based on n-grams achieve comparable performance, which suggests that, in dialogue, 'local' features can be indicative of the language of speakers from different age groups. We showed this to be the case, with both lexical and stylistic cues being informative to these models in this task.

While we performed the classification task at the level of single dialogue utterances, future work may take into account larger dialogue fragments, such as the entire dialogue or a fixed number of turns. This would make the setup more comparable to discourse, but would require making experimental choices and dealing with extra computational challenges. Moreover, it could be tested whether the language used by a speaker is equally discriminative when talking to a same-age (this work) or a different-age interlocutor.

Finally, we believe our findings could inform future work on developing adaptive conversational systems. Since consistent language style differences were found between age groups (for example, at the level of exclamatives and acknowledgments), systems whose language generation capabilities aim to be consistent with a given age group should therefore reproduce these patterns. This could be achieved, for example, by embedding one or more discriminative modules that control the generation of a system's output, which could lead to better, more natural interactions between human speakers and a conversational system.

Acknowledgements

This work received funding from the University of Amsterdam's Research Priority Area *Human(e) AI* and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

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, June. Association for Computational Linguistics.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark, September. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ming Li, Kyu J Han, and Shrikanth Narayanan. 2013. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language*, 27(1):151–167.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528.

R Love, C Dembry, A Hardie, V Brezina, and T McEnery. 2017. The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *In International Journal of Corpus Linguistics*, 22(3):319–344.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020. Plug-and-play conversational models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2422–2433, Online, November. Association for Computational Linguistics.

Michael McTear. 2020. Conversational AI: Dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies*, 13(3):1–251.

James W Pennebaker and Lori D Stone. 2003. Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85(2):291–301.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Margot J van der Goot and Tyler Pilgrim. 2019. Exploring age differences in motivations for and acceptance of chatbot communication in a customer service context. In *International Workshop on Chatbot Research and Design*, pages 173–186. Springer.

Maria Wolters, Ravichander Vipperla, and Steve Renals. 2009. Age recognition for spoken dialogue systems: Do we need it? In *Tenth Annual Conference of the International Speech Communication Association (Interspeech)*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online, November. Association for Computational Linguistics.

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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *CoRR*, abs/1901.09672.