The Multimodal Correction Detection Problem

Extended Abstract

Amos Azaria
Computer Science Department, Ariel University
Ariel, Israel
amos.azaria@ariel.ac.il

Keren Nivasch
Computer Science Department, Ariel University
Ariel, Israel
kerenni@ariel.ac.il

ABSTRACT

In order for socially aware agents to be truly useful, they should have abilities associated with human intelligence, such as the ability to detect their own mistakes from user reactions. This is an instance of *implicit feedback*.

In this work we address the problem of detecting an agent's mistakes by identifying when the user tries to correct the agent. We refer to this problem as the Correction Detection task. We use a multimodal approach, using both the voice (acoustics and non-verbal sounds) as well as the transcript of the user's spoken commands.

CCS CONCEPTS

• **Computing methodologies → Speech recognition**;

KEYWORDS

Human-agent interaction; Correction detection; Implicit feedback; Multimodal deep learning architecture; Socially aware personal assistant

ACM Reference Format:
Amos Azaria and Keren Nivasch. 2019. The Multimodal Correction Detection Problem. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Intelligent agents that are able to interact with users using natural language are becoming increasingly common. Popular operating systems now come with built-in virtual assistants, such as Siri for Apple's MacOS and iOS, and Cortana for Microsoft's Windows. As another example, Amazon's Echo speakers include the Alexa virtual assistant. However, these assistants do not learn from their own mistakes, in contrast to real human assistants.

When humans interact with one another, it often happens that one person misunderstands the other. This person might then realize that she made a mistake by the other person's reaction. As a consequence, she will not only correct her mistake, but she will also learn for the future what the other person's intentions were in such a situation. For example, when a manager tells her human assistant "I would like to promote Mary", the assistant might reply "I sent an email to Mary with the subject 'You're promoted'." Then the manager might reply "I would like to set a meeting to promote her". The human assistant will then probably recall the email and schedule a meeting with Mary for the promotion. The next time the manager tells the assistant she would like to promote someone, the assistant will remember to set up a promotion meeting.

If a socially aware agent had the ability to detect its own mistakes from user reactions, the agent would be much more useful. This is an instance of *implicit feedback*, which is the gathering of information from users' behavior, as they go along normally using the agent.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

2 PROBLEM FORMULATION

In this work we address the problem of detecting an agent's mistakes by identifying when the user tries to correct the agent. We refer to this problem as the Correction Detection task.

A social agent with the ability to detect user corrections might be able to fix some of the mistakes it makes. For example, suppose a user says "create an email for Tom", and the agent creates a new email and sets the address to Tom's address. Then the user says "create an email and set the subject to for Tom". Then the agent might erase the email it created and create a new email in which the subject is set to "For Tom".

In addition, an agent might learn for the future what a particular user means when giving a certain kind of request. In the above example, if later on the user says "create an email for Nancy", then the agent will create a new email and set the subject to "For Nancy".

In order to study this problem, we worked on a data-set in which each pair of consecutive commands has one of three possible labels: "new command" if the user was satisfied with the agent's action to the previous command and issued a new command; "command correction" if the user was not satisfied with the agent's action and tried to correct it; and "ASR correction" if the first command was not carried out properly due to wrong transcription by the Automatic Speech Recognition (ASR) system (for example, "set subject to Johnny" instead of "set subject to join me").

Assume a a dataset of size n consisting of multiple users interacting with a personal assistant agent. Let $C = \{c_1, c_2, ..., c_n\}$ be a set of commands a given to a personal agent. Each of the commands, c, is composed of a transcript of the command, c^t, and the command voice, c^v. Let $t(c_i, c_j)$ be a function that associates commands c_i and c_j with a type in $\{new, asr, cc\}$, where *new* denotes no relation between two commands (that is, the c_j is a new command), *asr* denotes that the c_j was given in order to correct a malperformance of the transcription performed by the agent, and *cc* denotes that the c_j, is an attempt of the user to refine and correct the c_i. In this paper we focus on the consecutive multimodal correction detection problem, in which, for each command, c_i, the value of $t(c_i, c_{i+1})$ must be determined.

It is important to separate command corrections from ASR corrections, because the actions to be taken by the agent are very

Command 1	Agent Response	Command 2	Label
create new email	The "recipient list" field in "outgoing email" was set to: "tom@gmail.com". What shall I do next (when executing: "tell Tom I'm running late")?	set recipient to Tom at gmail.com	"new command"
create a contact	Concept "contact" was defined successfully. Please add fields to it.	a contact has an email address	"new command"
I'm running late	Sorry, but I don't understand. Would you like to teach me (say yes or just ignore)?	tell Tom I'm running late	"command correction"
how long can a whale stay underwater	Sorry, but no answer was found.	how long can whale wait underwater	"command correction"
play the email	{command:playYouTube, item:qXlgOX95Q0U}	read email	"command correction"
who was the present of the United States	Tuesday, June 20, 2017	who was the president of the United States	"ASR correction"
in for Mariam late	Sorry, but I don't understand. Would you like to teach me (say yes or just ignore)?	inform Mary I'm late	"ASR correction"

Table 1: Data-set examples

different. With an ASR correction, the agent should adjust the ASR component and improve it, so that it does not fail next time. However, when dealing with a command correction, the agent should undo the previous command, and execute the learning process, as it has implicitly learned another way to say the second command.

Our approach is multimodal as it uses both the voice (acoustics and non-verbal sounds) as well as the transcript of the user's spoken commands. The voice input can hold many important cues, such as tone, speed, or emphasis on certain words. Further, voice input can be especially useful in cases where the wrong command was executed due to a fault in the ASR.

3 RELATED WORK

Implicit feedback has received a great deal of attention [10, 13, 19]. In particular, search engines can use implicit feedback in order to improve the ranking of search results [1, 11, 12, 20]. The act of down-ranking one search result and up-ranking another can be considered a correction performed by the search engine in response to the user's behavior.

Paraphrase detection is the task of deciding whether two given sentences have the same meaning even though they use different words [5, 7–9, 14, 17, 21, 23, 24]. Paraphrase detection is closely related to our Correction Detection problem. Indeed, a user might try to correct an agent by repeating the previous command in slightly different words. For example, the user might give the command "remove the contact Tom" and the agent might not understand or not perform it correctly. Then the user might try again in different words by saying "delete the contact named Tom".

However, there are some differences between paraphrase detection and the Correction Detection task. The second command might constitute a correction of the first, even though it has a slightly different meaning: The two commands might differ in proper names (e.g. Tom vs. John) or in numerical quantities, and the user's tone of voice might indicate that he got confused in the first command. Furthermore, in our task the order of the commands might be significant. For example, the agent might understand the word "create" but not the word "compose". Hence, the order between the commands "create an email for Tom" and "compose an email for Tom" is very significant.

Multimodal deep learning has been applied to tasks such as speech recognition, speech synthesis, emotion and affect detection, media description, and multimedia retrieval [4, 15, 18, 22]. As far as we know, this is the first work to apply multimodal voice and transcript deep learning for Correction Detection.

4 THE DATA-SET

We collected a set of real interactions that users had while experimenting with the social agent LIA [3, 6, 16]. Our data-set contains a series of spoken commands given to LIA by different users. For each command we have the original voice file and the written transcript produced by the ASR. We manually labeled each pair of consecutive commands according to whether the second one is a correction of the first. See Table 1 for some examples.

5 ACKNOWLEDGMENTS

This work was a part of the InMind project for the creation of a smart personal assistant [2].

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 19–26.

[2] Amos Azaria and Jason Hong. 2016. Recommender systems with personality. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 207–210.

[3] Amos Azaria, Jayant Krishnamurthy, and Tom M. Mitchell. 2016. Instructable Intelligent Personal Agent. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016*. AAAI, Phoenix, Arizona, USA, 2681–2689. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12383

[4] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal Machine Learning: A Survey and Taxonomy. *CoRR* abs/1705.09406 (2017). arXiv:1705.09406 http://arxiv.org/abs/1705.09406

[5] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating Sentences from a Continuous Space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 10–21.

[6] Merav Chkroun and Amos Azaria. 2019. LIA: A Virtual Assistant that Can Be Taught New Commands by Speech. *International Journal of Human–Computer Interaction* (2019), 1–12.

[7] Dipanjan Das and Noah A Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. 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: Volume 1-Volume 1*. Association for Computational Linguistics, 468–476.

[8] Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 350.

[9] Andrew Finch, Young-Sook Hwang, and Eiichiro Sumita. 2005. Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

[10] Dietmar Jannach, Lukas Lerche, and Markus Zanker. 2018. Recommending Based on Implicit Feedback. In *Social Information Access - Systems and Technologies*. 510–569. https://doi.org/10.1007/978-3-319-90092-6_14

[11] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, Vol. 51. Acm, 4–11.

[12] Thorsten Joachims and Filip Radlinski. 2007. Search Engines that Learn from Implicit Feedback. *IEEE Computer* 40, 8 (2007), 34–40. https://doi.org/10.1109/MC.2007.289

[13] Diane Kelly and Jaime Teevan. 2003. Implicit feedback for inferring user preference: a bibliography. In *ACM SIGIR Forum*, Vol. 37. ACM, 18–28.

[14] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*. 3294–3302.

[15] Toby Jia-Jun Li, Amos Azaria, and Brad A Myers. 2017. SUGILITE: creating multimodal smartphone automation by demonstration. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6038–6049.

[16] Toby Jia-Jun Li, Igor Labutov, Brad A Myers, Amos Azaria, Alexander I Rudnicky, and Tom M Mitchell. 2018. An End User Development Approach for Failure Handling in Goal-oriented Conversational Agents. *Studies in Conversational UX Design* (2018).

[17] Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 182–190.

[18] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 689–696.

[19] Douglas W Oard and Jinmook Kim. 2001. Modeling information content using observable behavior. In *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology*. ASIS&T, USA, 38–45.

[20] Filip Radlinski and Thorsten Joachims. 2005. Query chains: learning to rank from implicit feedback. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 239–248.

[21] Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*. 801–809.

[22] Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep Boltzmann machines. In *Advances in neural information processing systems*. 2222–2230.

[23] Stephen Wan, Mark Dras, Robert Dale, and Cécile Paris. 2006. Using dependency-based features to take the 'para-farce' out of paraphrase. In *Proceedings of the Australasian Language Technology Workshop 2006*. 131–138.

[24] Yuhuai Wu, Saizheng Zhang, Ying Zhang, Yoshua Bengio, and Ruslan R Salakhutdinov. 2016. On multiplicative integration with recurrent neural networks. In *Advances in neural information processing systems*. 2856–2864.