Generating Subjective Responses to Opinionated Articles in Social Media: An Agenda-Driven Architecture and a Turing-Like Test

Tomer Cagan
School of Computer Science
The Interdisciplinary Center
Herzeliya, Israel
cagan.tomer@idc.ac.il

Stefan L. Frank
Centre for Language Studies
Radboud University
Nijmegen, The Netherlands
s.frank@let.ru.nl

Reut Tsarfaty
Mathematics and Computer Science
Weizmann Institute of Science
Rehovot, Israel
tsarfaty@weizmann.ac.il

Natural language traffic in social media enjoys vast monitoring and analysis efforts. However, the question whether computer systems can generate such content has been only sparsely attended to. This paper presents an end-to-end architecture for generating subjective responses to online articles. We aim to generate responses that promote specific users' agendas, ones that may be different than the opinion reflected in the article. Our generation system integrates users' agendas, documents' topics, sentiment analysis and a knowledge graph, alongside a template-based surface realizer. We present a novel empirical evaluation method for quantifying the human-likeness and relevance of the generated responses in a ground-sourced, Turing test-like setting. We empirically show that including world knowledge in the input increases the generated responses' human-likeness, while not affecting perceived relevance.[1]

Motivation and Background

As many of our day-to-day activities move online (Viswanath et al., 2009), the importance of social media interactions to businesses (Qualman, 2012; Haenlein and Kaplan, 2009) and governments (Howard et al., 2011; Lamer, 2012) vastly increases. Therefore, natural language traffic in social media (blogs, microblogs, talkbacks) now enjoys vast monitoring and analysis efforts in such organizations. The increased importance of online communication has also led to many research advances that pertain to the analysis of social-media interactions: subjectivity and sentiment analysis (Davidov et al., 2010), opinion mining (Mishne, 2006), affectiveness of online texts (Danescu-Niculescu-Mizil et al., 2009) and many more. In contrast to research on such *analysis* efforts, the question whether computer systems can *generate* online content to effectively interact with humans has been only sparsely attended to.[2]

Research Question

Can a computer generate a fluent, relevant, and human-like response which effectively engages readers and clearly serves the responder's communicative goal? The present paper addresses this problem of generating novel, subjective, responses to online opinionated articles on behalf of an interested agent. We propose a formal model, an end-to-end implemented architecture, and an empirical evaluation method for a system that generates such responses. The generation process takes into account the user's agenda, the document's topics, sentiment status and, optionally, a knowledge graph. In addressing the question how such responses may be faithfully evaluated, we develop a Turing test-like method for quantifying the human-likeness and relevance of computer-generated responses in online settings.

The Solution

In social media, natural language utterances are often employed as a communicative device serving a communicative goal, such as promoting the user's disposition towards some topic. To define this communicative goal, we define a *user agenda* as a set of topics associated with the user's sentiment. A triggering event for generating an utterance that serves such goal may be a new online *document* which conveys an authors sentiment toward some topics. When the agenda and the document content overlap, a response generation is triggered by our system.

[1] This work is now published as Cagan et al. (2014)

[2] The only study we are aware of is Ritter et al. (2011), which uses a machine translation engine to generate responses to tweets. This work differs from our own in that it does not generate novel subjective responses, but rather, provides one-size-fits-all mechanism.

Formally, assuming that A is a set of agendas, D a set of documents and S a set of valid English sentences, we wish to implement the following: $f_{\text{response}} : D \times A \rightarrow S$, where S is an English sentence expressing the responder's beliefs or sentiments towards the document topics, and relative to that of the author. To implement this function we define a composite process containing two phases: (a) an analysis phase $p : D \rightarrow C$, and (b) a generation phase $g : C \times A \rightarrow S$.

The goal of p is to extract the document's topic(s) and related sentiments (henceforth, a *content element*) and yield a set of content elements represented as $c \in C$). The generation phase takes as input the content elements extracted from the document as well as the content elements defined in the user's agenda and generates a response based on their intersection. The implementation of p relies on a trained topic model (Papadimitriou et al., 1998; Hofmann, 1999; Blei et al., 2003) with an associated sentiment value, $sentiment_t \in [-n..n]$, defined for each document or user agenda. The implementation of g employs a template-based generation approach, as in Reiter and Dale (1997), Van Deemter et al. (2005).

The design of our template reflects the three Gricean maxims of communication (Grice, 1967): *quantity* (responses are brief and concise), *relation* (responses directly address the documents content) and *quality* (responses express responders beliefs, sentiments, or dispositions towards the topic(s)). We enforce these maxims through the templates' design: the responses length and density is controlled by the number substitution slots in the templates (quantity), templates directly incorporate user/document topics and sentiments,(relation). and users opinions, perceived as their respective truth, define the responses relation to the document content (quality).

Empirical Evaluation

Based on our architecture, we implemented and tested two systems. A baseline system as defined above, and an additional variant that also includes a knowledge-base that can be used to expand the response with a sentence on topics related to that of the document. Due to the large space of output possibilities there is no gold standard or ground-truth to compare our generated responses to, and we resort to human evaluation akin to the well-known *Turing test* (Turing, 1950). In our evaluation, we ask human participants to evaluate our system output as well as real human responses for the same articles snippets. We conducted two online surveys (using Amazon Mechanical Turk - `www.mturk.com`) in which we asked the participants to rate the human-likeness and relevance of the human and computer responses. In all cases, we considered online articles on mobile devices, and simulated responses for a range of possible users agendas. Some response are generated with the addition of a knowledge-base, and others without.

Results

Our generated responses achieve a computer-likeness rating higher than that of human responses (4.32 rating for the system and 3.33 for human responses), indicating that our ultimate goal is yet to be reached. In terms of relevance, our response scored 4.52 while human responses was at 4.85, indicating that in terms of relevance, our generated responses is roughly at the same level as human responses. We additionally investigated, using regression analysis, what factors makes responses more human-like. Our results show that responses generated using world knowledge are regarded as more human-like than those that rely on topic, sentiment and agenda only – whereas the use of world knowledge does not affect perceived relevance. We also identified a learning effect of participants, getting better at identifying the computer responses over time, which we attribute to the repetitiveness in the use of our various templates.

Conclusion and Future Work

Our evaluation exposed several strength and weaknesses of the models, which we aim to further investigate and improve on in future work. Firstly, we aim to use our empirical evaluation method to study online responses more comprehensively, towards identifying common linguistic characteristics. We then plan to use these linguistic characteristics for devising a more general grammar-based generation engine replacing our templates, combatting the learning effect by adding more variance. On a different note, we plan to explore the use of a wide-scale knowledge base, such as Freebase (Bollacker et al., 2008) in order to expand our output domain and make responses more human-like, more diverse, and ultimately also more interesting.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, New York, NY, USA. ACM.

Tomer Cagan, L. Stefan Frank, and Reut Tsarfaty, 2014. *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, chapter Generating Subjective Responses to Opinionated Articles in Social Media: An Agenda-Driven Architecture and a Turing-Like Test, pages 58–67. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How opinions are received by online communities: A case study on amazon.com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 141–150, New York, NY, USA. ACM.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 241–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

H. P. Grice. 1967. Logic and conversation. In H. P. Grice, editor, *Studies in the ways of words*, pages 22–40. Harvard University Press.

Michael Haenlein and Andreas M. Kaplan. 2009. Flagship brand stores within virtual worlds: The impact of virtual store exposure on real-life attitude toward the brand and purchase intent. *Recherche et Applications en Marketing (English Edition)*, 24(3):57–79.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM.

Philip N. Howard, Aiden Duffy, Deen Freelon, Muzammil Hussain, Will Mari, and Marwa Mazaid. 2011. Opening closed regimes: What was the role of social media during the Arab spring? *Project on Information Technology and Political Islam.*

Wiebke Lamer. 2012. Twitter and tyrants: New media and its effects on sovereignty in the Middle East. *Arab Media and Society*.

Gilad Mishne. 2006. Multiple ranking strategies for opinion retrieval in blogs. In *Proceedings of the 15th Text Retrieval Conference*.

Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. 1998. Latent Semantic Indexing: A probabilistic analysis. In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, PODS '98, pages 159–168, New York, NY, USA. ACM.

Erik Qualman. 2012. *Socialnomics: How social media transforms the way we live and do business*. John Wiley & Sons, Hoboken, NJ, USA, 2nd edition.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Nat. Lang. Eng.*, 3(1):57–87.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 583–593, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alan M. Turing. 1950. Computing machinery and intelligence. *Mind*, LIX:433–460.

Kees Van Deemter, Emiel Krahmer, and Mariët Theune. 2005. Real versus template-based natural language generation: A false opposition? *Comput. Linguist.*, 31(1):15–24.

Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. 2009. On the evolution of user interaction in Facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, WOSN '09, pages 37–42, New York, NY, USA. ACM.