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

Abstract

Natural language traffic in social media (blogs, microblogs, talkbacks) enjoys vast monitoring and *analysis* efforts. However, the question whether computer systems can *generate* such content in order to effectively interact with humans has been only sparsely attended to. This paper presents an architecture for generating subjective responses to opinionated articles based on users' agenda, documents' topics, sentiments and a knowledge graph. We present an empirical evaluation method for quantifying the human-likeness and relevance of the generated responses. We show that responses generated using world knowledge in the input 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.

1 Introduction

Digital media, user-generated content and social networks enable effective human interaction; so much so that much of our day-to-day interaction is conducted online (Viswanath et al., 2009). Interaction in social media fundamentally changes the way businesses and consumers behave (Qualman, 2012), can be instrumental to the success of individuals and businesses (Haenlein and Kaplan, 2009), and even affects the stability of political regimes (Howard et al., 2011; Lamer, 2012). These facts force organizations (businesses, governments, and non-profit organizations) to be constantly involved in the monitoring of, and the interaction with, human agents in digital environments (Langheinrich and Karjoth, 2011).

Automatic analysis of user-generated online content benefits from extensive research and commercial opportunities. In natural language processing, there is ample research on the analysis of subjectivity and sentiment of content in social media. The development of tools for sentiment analysis (Davidov et al., 2010), mood aggregation (Agichtein et al., 2008), opinion mining (Mishne, 2006), and many more, now enjoys wide interest and exposure, as is also evident by the many workshops and dedicated tracks at ACL venues.[1] Methods are also developed for the analysis of political texts (O'Connor et al., 2010; O'Connor et al., 2013) and for text-driven forecasting based on these data (Yano et al., 2009). A related strand of research uses computational methods to find out what kind of published utterances are influential, and how they affect linguistic communities (Danescu-Niculescu-Mizil et al., 2009). Such work complements, and contributes to, studies from sociology and sociolinguistics that aim to delineate the process of generating meaningful responses (e.g., Amabile (1981)).

In contrast to these analysis efforts, the topic of *generating* responses to content in social media is only sparsely explored. Commercially, there is movement towards online response automation (Owyang, 2012; Mah, 2012).[2] Research on user interfaces is trying to move away from script-based interaction towards the development of chat bots that attempt natural human-like interaction (Mori et al., 2003; Feng et al., 2006). However, these chat bots are typically designed to provide an automated one-size-fits-all type of interaction.

A study by Ritter et al. (2011) addresses the generation of responses to natural language tweets in a data-driven setup. It applies a machine-translation approach to response generation, where moods and sentiments already ex-

[1] E.g., the ACL series LASM http://tinyurl.com/ludyrkz; WASSA http://tinyurl.com/kjjdhax.

[2] There is a general debate on the efficiency of automated tools (Nall, 2013) and whether such tools are desirable in social media (McConnell (2012); responses to Owyang (2012)).

Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media, pages 58–67,
Baltimore, Maryland USA, 27 June 2014.

pressed in the past are replicated or reused. A recent study by Hasegawa et al. (2013) modifies Ritter's approach to produce responses that elicit an emotion from the addressee. Yet, these responses do not target particular topics and are not driven by a user agenda.

The present paper addresses the problem of generating novel, subjective, responses to online opinionated articles. We formally define the document-to-response mapping problem and suggest an end-to-end system to solve it. Our system integrates a range of NLP and NLG technologies (including topic models, sentiment analysis, and the integration of a knowledge graph) to design a flexible generation mechanism that allows us to vary the information in the input to the generation procedure. We then use a Turing-inspired test to study the different factors that contribute to the perceived human-likeness and relevance of the generated responses, and show how the perception of responses depends on external knowledge and the expressed sentiment.

The remainder of this paper is organized as follows. The next section presents our proposal: Section 2.1 describes our approach, Section 2.2 formalizes the proposal, and Section 2.3 presents our end-to-end architecture. This is followed by our evaluation method and empirical results in Section 3. We discuss related and future work in Section 4, and in Section 5 we conclude.

2 The Proposal: Generating Subjective Responses

2.1 Our Approach

Natural language is, above all, a communicative device that we employ to achieve certain goals. In social media, the driving force behind generating responses is a responder's disposition towards some topic. This topic could be a political campaign or a candidate, a product, or some abstract idea, which the responder has a motive to promote. Let us call this goal our user's *agenda*.

User response generation, like any other natural language utterance generation, is triggered by a certain event that is related to the communicative goal. In a social media setting, this event is often a new online *document*. The document and the agenda thus form the input to our generation system. Each document and each agenda contain (possibly many) topics, each of which is associated with a (positive or negative) sentiment. Document sentiments are attributed to the author, whereas agenda sentiments are attributed to the user (henceforth: the responder).

For each non-empty intersection of the topics in the document and in the agenda, our response-generation system aims to generate utterances that are fluent, human-like, and effectively engage readers. The generation is based on three assumptions, roughly reflecting the Gricean maxims of cooperative interaction (Grice, 1967). Online user responses should then be:

- *Economic* (Maxim of Quantity): Responses are brief and concise;
- *Relevant* (Maxim of Relation): Responses directly address the documents' content.
- *Opinionated* (Maxim of Quality): Responses express responders beliefs, sentiments, or dispositions towards the topic(s).

2.2 The Formal Model

Let D be a set of documents and let A be a set of user agendas as we define shortly. Let S be a set of English sentences over a finite vocabulary $S = \Sigma^*$. Our system implements a function that maps each $\langle document, agenda \rangle$ pair to a natural language response sentence $s \in S$.

$$f_{\text{response}} : D \times A \rightarrow S$$

Response generation takes place in two phases, roughly corresponding to macro and micro planning in Reiter and Dale (1997):

- Macro Planning (below, the *analysis* phase): What are we going to talk about?
- Micro Planning (below, the *generation* phase): How are we going to say it?

The analysis function $p : D \rightarrow C$ maps a document to a subjective representation of its content.[3] The generation function $g : C \times A \rightarrow S$ intersects the content elements in the document and in the user agenda, and generates a response based on the content of the intersection. All in all, our system implements a composition of the analysis and the generation functions:

$$f_{\text{response}}(d, a) = g(p(d), a) = s$$

[3] A content element may conceivably encompass a topic, its sentiment, its objectivity, its evidentiality, its perceived truthfulness, and so on. In this paper we focus on topic and sentiment, and leave the rest for future research.

Each content element $c \in C$ or an agenda item $a \in A$ is composed of a topic t associated with a sentiment value $sentiment_t \in [-n..n]$ that signifies the (negative or positive) disposition of the document's author (if $c \in C$) or the user's agenda (if $a \in A$) towards the topic. We assume here that a topic is simply a bag of words from our vocabulary Σ. Thus, we have the following:

$$A, C \subseteq \mathcal{P}(\Sigma) \times [-n..n]$$

Our generation component accepts the result of the intersection as input and relies on a template-based grammar and a set of functions for generating referring expressions in order to construct the output. To make the responses *economic*, we limit the content of a response to one statement about the document or its author, followed by a statement on the relevant topic. To make the response *relevant*, the templates that generate the response make use of topics in the intersection of the document and the agenda. To make the response *opinionated*, the sentiment of the response depends on the (mis)match between the sentiment values for the topic in the document and in the agenda. Concretely, the response is positive if the sentiments for the topic in the document and agenda are the same (both positive or both negative) and it is negative otherwise.

We suggest two variants of the generation function g. The basic variant implements the baseline function defined above:

$$g_{\text{base}}(c, a) = s$$

$$c \in C, a \in A, s \in \Sigma^*$$

For the other variant we define a knowledge base (KB) as a directed graph in which words $w \in \Sigma$ from the topic models correspond to nodes in the graph, and relations $r \in R$ between the words are predicates that hold in the real world. Our second generation function now becomes:

$$g_{\text{kb}}(c, a, \mathit{KB}) = s$$

$$\mathit{KB} \subseteq \{(w_i, r, w_j) | w_i, w_j \in \Sigma, r \in R\}$$

with $c \in C, a \in A, s \in \Sigma^*$ as defined in g_{base} above.

2.3 The Architecture

The system architecture from a bird's eye view is presented in Figure 1. In a nutshell, a document enters the analysis phase, where topic inference and sentiment scoring take place, resulting in ⟨topic, sentiment⟩-pairs. During the subsequent generation phase, these are intersected with the ⟨topic, sentiment⟩-pairs in the user agenda. This intersection, possibly augmented with a knowledge graph, forms the input for a template-based generation component.

Analysis phase For the task of inferring the topics of the document we use topic modeling: a probabilistic generative modeling technique that allows for the discovery of abstract topics over a large body of documents (Papadimitriou et al., 1998; Hofmann, 1999; Blei et al., 2003). Specifically, we use topic modeling based on *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003; Blei, 2012). Given a new document and a trained model, the inference method provides a weighted mix of topics for that document, where each topic is represented as a vector containing keywords associated with probabilities. For training the topic model and inferring the topics in new documents we use *Gensim* (Rehurek and Sojka, 2010), a fast and easy-to-use implementation of LDA.

Next, we wish to infer the sentiment that is expressed in the text with relation to the topic(s) identified in the document. We use the semantic/lexical method as implemented in Kathuria (2012). We rely on a WSD sentiment classifier that uses the SentiWordNet (Baccianella et al., 2010) database and calculates the positivity and negativity scores of a document based on the positivity and negativity of individual words. The result of the sentiment analysis is a pair of values, indicating the positive and negative sentiments of the document-based scores for individual words. We use the larger of these two values as the sentiment value for the whole document.[4]

Generation phase Our generation function first intersects the set of topics in the document and the set of topics in the agenda in order to discover relevant topics to which the system would generate responses. A response may in principle integrate content from a range of topics in the topic model distribution, but, for the sake of generating concise responses, in the current implementation we focus on the single most prevalent, topic. We pick the highest scoring word of the highest scoring topic, and intersect it with topics in the agenda. The system generates a response based on the identified

[4]Clearly, this is a simplifying assumption. We discuss this assumption further in Section 4.

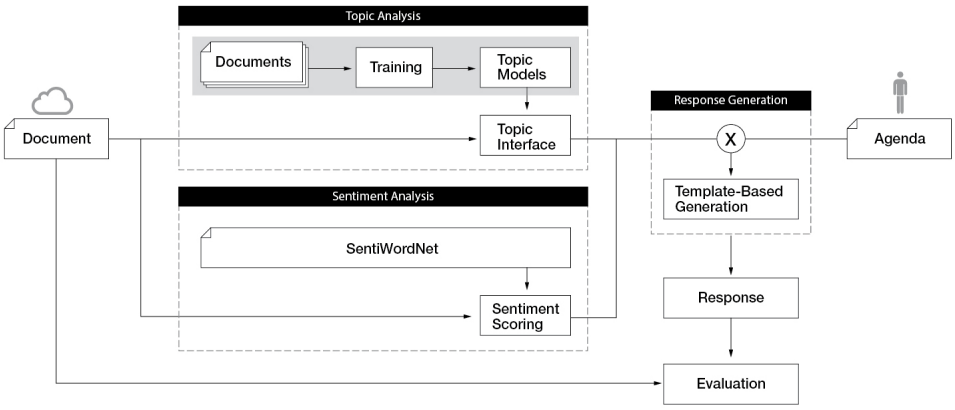

Figure 1: The system architecture from a bird's eye view. Components on gray background are executed offline.

topic, the sentiment for the topic in the document, and the sentiment for that topic in the user agenda.

The generation component relies on a template-based approach similar to Reiter and Dale (1997) and Van Deemter et al. (2005). Templates are essentially subtrees with leaves that are placeholders for other templates or for functions generating referring expressions (Theune et al., 2001). These functions receive (relevant parts of) the input and emit the sequence of fine-grained part-of-speech (POS) tags that realizes the relevant referring expression. The POS tags in the resulting sequences are ultimately place holders for words from a lexicon Σ. In order to generate a variety of expression forms — nouns, adjectives and verbs — these items are selected randomly from a fine-grained lexicon we defined. The sentiment (positive or negative) is expressed in a similar fashion via templates and randomly selected lexical entries for the POS slots, after calculating the overall sentiment for the intersection as stated above. Our generation implementation is based on SimpleNLG (Gatt and Reiter, 2009) which is a surface realizer API that allows us to create the desired templates and functions, and aggregates content into coherent sentences. The templates and functions that we defined are depicted in Figure 2.

In addition, we handcrafted a simple knowledge graph (termed here KB) containing the words in a set of pre-defined user agendas. Table 1 shows a snippet of the constructed knowledge graph. The knowledge graph can be used to expand the response in the following fashion: The topic of the response is a node in the KB. We randomly select one of its outgoing edges for creating a related

Source	Relation	Target
Apple	CompetesWith	Samsung
Apple	CompetesWith	Google
Apple	Creates	iOS

Table 1: A knowledge graph snippet.

statement that has the target node of this relation as its subject. The related sentence generation uses the same template-based mechanism as before. In principle, this process may be repeated any number of times and express larger parts of the KB. Here we only add one single knowledge-base relation per response, to keep the responses concise.

3 Evaluation

We set out to evaluate how computer-generated responses compare to human responses in their perceived *human-likeness* and *relevance*. More in particular, we compare different system variants in order to investigate what makes responses seem more human-like or relevant.

3.1 Materials

Our empirical evaluation is restricted to topics related to mobile telephones, specifically Apple's iPhone and devices based on the Android operating system. We collected 300 articles from leading technology sites in the domain to train the topic models on, settling on 10 topics models. Next, we generated a set of user agendas referring to the same 10 topics. Each agenda is represented by a single keyword from a topic model distribution and a sentiment value $sentiment_t \in \{-8, -4, 0, 4, 8\}$. Finally, we selected 10 new articles from similar sites and generated a pool of

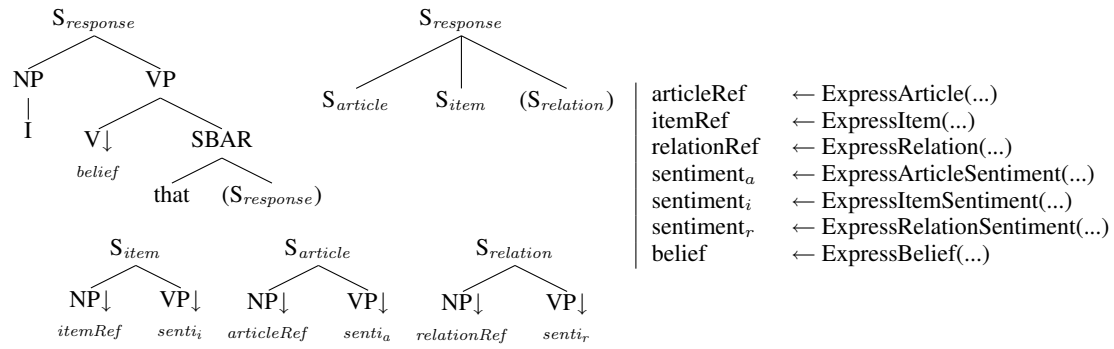

Figure 2: Template-based response generation. The templates are on the left. The Express* functions on the right uses regular expressions over the arguments and vocabulary items from a closed lexicon.

1000 responses for each, comprising 100 unique responses for each combination of $sentiment_t$ and system variant (i.e., with or without a knowledge base). Table 2 presents an example response for each such combination. In addition, we randomly collected 5 to 10 real, short or medium-length, online human responses for each article.

3.2 Surveys

We collected evaluation data via two online surveys on Amazon Mechanical Turk (`www.mturk.com`). In Survey 1, participants judged whether responses to articles were written by human or computer, akin to (a simplified version of) the Turing test (Turing, 1950). In Survey 2, responses were rated on their relevance to the article, in effect testing whether they abide by the Gricean Maxim of Relation. This is comparable to the study by Ritter et al. (2011) where people judged which of two responses was 'best'.

Each survey comprises 10 randomly ordered trials, corresponding to the 10 selected articles. First, the participant was presented with a snippet from the article. When clicking a button, the text was removed and its presentation duration recorded. Next, a multiple-choice question asked about the snippet's topic. Data on a trial was discarded from analysis if the participant answered incorrectly or if the snippet was presented for less than 10 msec per character; we took these to be cases where the snippet was not properly read. Next, the participant was shown a randomly ordered list of responses to the article.

In Survey 1, four responses were presented for each article: three randomly selected from the pool of human responses to that article and one generated by our system. The task was to categorize each response on a 7-point scale with labels 'Certainly human/computer', 'Probably human/computer', 'Maybe human/computer' and 'Unsure'. In Survey 2, five responses were presented: three human responses and two computer-generated. The task was to rate the responses' relevance on a 7-point scale labeled 'Completely (not) relevant', 'Mostly (not) relevant', 'Somewhat (not) relevant', and 'Unsure'. As a control condition, one of the human responses and one of the computer responses were actually taken from another article than the one just presented. In both surveys, the computer-generated responses presented to each participant were balanced across sentiment levels and generation functions (g_{base} and g_{kb}). After completing the 10 trials, participants provided basic demographic information, including native language. Data from non-native English speakers was discarded. Surveys 1 and 2 were completed by 62 and 60 native speakers, respectively.

3.3 Analysis and Results

Survey 1: Computer-Likeness Rating. Table 3 shows the mean 'computer-likeness'-ratings from 1 ('Certainly human') to 7 ('Certainly computer') for each response category. Clearly, the human responses are rated as more human-like than the computer-generated ones: our model did not generally mislead the participants. This may be due to the template-based response structure: over the course of the survey, human raters are likely to notice this structure and infer that such responses are computer-generated. To investigate whether such learning indeed occurs, a linear mixed-effects model was fitted, with predictor variables IS_COMP ($+1$:computer-generated, -1:human responses), POS (position of the trial in the survey, 0 to 9), and the interaction between the two. Table 4

Sent.	KB	Response
−8	No	Android is horrendous so I think that the writer is completely correct!!!
	Yes	Apple is horrendous so I feel that the author is not really right!!! iOS is horrendous as well.
−4	No	I think that the writer is mistaken because apple actually is unexceptional.
	Yes	I think that the author is wrong because Nokia is mediocre. Apple on the other hand is pretty good ...
0	No	The text is accurate. Apple is okay.
	Yes	Galaxy is okay so I think that the content is accurate. All-in-all samsung makes fantastic gadgets.
4	No	Android is pretty good so I feel that the author is right.
	Yes	Nokia is nice. The article is precise. Samsung on the other hand is fabulous...
8	No	Galaxy is great!!! The text is completely precise.
	Yes	Galaxy is awesome!!! The author is not completely correct. In fact I think that samsung makes awesome products.

Table 2: Responses generated by the system with or without a knowledge-base (KB), with different sentiment levels.

Response Type	Mean and CI
Human	3.33 ± 0.08
Computer (all)	4.49 ± 0.15
Computer (−KB)	4.66 ± 0.20
Computer (+KB)	4.32 ± 0.22

Table 3: Mean and 95% confidence interval of computer-likeness rating per response category. ±KB indicates whether g_{base} or g_{kb} was used.

Factor	b	t	$P(b < 0)$
(intercept)	3.590		
IS_COMP	0.193	2.11	0.015
POS	0.069	4.76	0.000
IS_COMP × POS	0.085	6.27	0.000

Table 4: Computer-likeness rating regression results, comparing human to computer responses.

presents, for each factor in the regression analysis, the coefficient b and its t-statistic. The coefficient equals the increase in computer-likeness rating for each unit increase in the predictor variable. The t-statistic is indicative of how much variance in the ratings is accounted for by the predictor. We also obtained a probability distribution over each coefficient by Markov Chain Monte Carlo sampling using the `R` package `lme4` version 0.99 (Bates, 2005). From each coefficient's distribution, we estimate the posterior probability that b is negative, which quantifies the reliability of the effect.

The positive b value for POS shows that responses drift towards the 'computer'-end of the scale. More importantly, a positive interaction with IS_COMP indicates that the difference between human and computer responses becomes more noticeable as the survey progresses — the participants did learn to identify computer-generated responses. However, the positive coefficient for IS_COMP means that even at the very first trial, computer responses are considered to be more computer-like than human responses.

Factors Affecting Human-Likeness. Our finding that the identifiability of computer-generated responses cannot be fully attributed to their repetitiveness, raises the question: What makes a such a response more human-like? The results provide several insights into this matter.

First, the mean scores in Table 3 suggest that including a knowledge base increases the responses' human-likeness. To further investigate this, we performed a separate regression analysis, using only the data on computer-generated responses. This analysis also included predictors KB ($+1$: knowledge base included, -1: otherwise), SENT ($sentiment_t$, from -8 to $+8$), absolute value of SENT, and the interaction between KB and POS. As can be seen in Table 5, there is no reliable interaction between KB and POS: the effect of including the KB on the human-likeness of responses remained constant over the course of the survey.

Furthermore, we see evidence that responses with a more positive sentiment are considered more computer-like. The (only weakly reliable) negative effect of the absolute value of sentiment suggests that more extreme sentiments are considered more human-like. Apparently, people count on computer responses to be mildly positive, whereas human responses are expected to be more extreme, and extremely negative in particular.

Survey 2: Relevance Rating. The mean relevance scores in Table 6 reveal that a response is rated as more relevant to a snippet if it was actually a response to that snippet, rather than to a different snippet. This reinforces our design choice

Factor	b	t	$P(b<0)$
(intercept)	4.022		
KB	−0.240	−2.13	0.987
POS	0.144	5.82	0.000
SENT	0.035	2.98	0.002
abs(SENT)	−0.041	−1.97	0.967
KB × POS	0.023	1.03	0.121

Table 5: Computer-likeness rating regression results, comparing systems with and without KB.

Factor	b	t	$P(b<0)$
(intercept)	3.861		
IS_COMP	−0.339	−7.10	1.000
SOURCE	0.824	16.80	0.000
IS_COMP × PRES	0.179	5.03	0.000

Table 7: Relevance ratings regression results, comparing human to computer responses.

to include input items referring specifically to the topic and sentiment of the author. However, human responses are considered more relevant than the computer-generated ones. This is confirmed by a reliably negative regression coefficient for IS_COMP (see regression results in Table 7).

The analysis included the binary factor SOURCE (+1 if the response came from the presented snippet, −1 if it came from a random article). We see a positive interaction between SOURCE and IS_COMP, indicating that presenting a response from a random article is more detrimental to relevance of computer-generated responses than that of the human responses. This is not surprising, as the computer-generated responses (unlike the human responses) always includes the article's topic.

When analyzing only data on computer-generated responses, and including predictors for agenda sentiment and for presence of the knowledge base, we see that including the KB does not affect response relevance (see Table 8). Also, there is no interaction between KB and SOURCE, that is, the effect of presenting a response from a different article does not differ between the models with and without the knowledge base. Possibly, responses are considered as more relevant if they have more positive sentiment, but the evidence for this is fairly weak.

Response Type	Source	Mean and CI
Human	this	4.85 ± 0.11
	other	3.56 ± 0.18
Computer (all)	this	4.52 ± 0.16
	other	2.52 ± 0.15
Computer (−KB)	this	4.53 ± 0.23
	other	2.46 ± 0.21
Computer (+KB)	this	4.51 ± 0.23
	other	2.58 ± 0.22

Table 6: Mean and 95% confidence interval of relevance rating per response category. 'Source' indicates whether the response is from the presented text snippet or a random other snippet. ±KB indicates whether g_{base} or g_{kb} was used.

Factor	b	t	$P(b<0)$
(intercept)	3.603		
KB	0.026	0.49	0.322
SOURCE	1.003	15.90	0.000
SENT	0.023	1.94	0.029
abs(SENT)	−0.017	−0.93	0.819
KB × SOURCE	−0.032	−0.61	0.731

Table 8: Relevance ratings regression results, comparing systems with and without KB.

4 Related and Future Work

In contrast to the vast amount of research on sentiment and topic analysis, as well as generation tasks in which the input is artificial or pre-defined, our system implements a full end-to-end cycle from natural language analysis to natural language generation with applications in social media and automated interaction in real-world settings.

The only two other studies on response generation in social media we know of are Ritter et al. (2011) and Hasegawa et al. (2013). Ritter's and Hasegawa's approaches differ from ours in their objective and their approach to generation. Specifically, Ritter's approach is based on machine translation, creating responses by directly re-using previous content. Their data-driven approach generates relevant, but not opinionated responses. In addition, both Ritter's and Hasegawa's systems respond to tweets, while our system analyzes and responds to complete articles. Hasegawa's approach is closer to ours in that it generates responses that are intended to elicit a specific emotion from the addressee. However, it still differs considerably in settings (dialogues versus online posting) and in the goal itself (eliciting emotion versus expressing opinion). Thus, we see these studies as complementary to ours in the realm of response generation in social media.

A natural contact point of our work with existing work in social media analysis is the investigation of how a change in the implementation of individual components (e.g., topic inference or sentiment scoring) would affect the result of the overall generation. In particular, it would be interesting to test whether a novel mechanism for joint inference of topic/sentiment distributions could lead to improvement in the human-likeness of the generated responses.

The syntactic and semantic means of expression that we use are based on bare bone templates and fine-grained POS tags (Theune et al., 2001). These may potentially be expanded with different ways to express subject/object relations, relations between phrases, polarity of sentences, and so on. Additional approaches to generation can factor in such aspects, e.g., the template-based methods in Becker (2002) and Narayan et al. (2011), or grammar based methods, as in DeVault et al. (2008). Using more sophisticated generation methods with a rich grammatical backbone may combat the sensitivity to computer-generated response patterns as acquired by our human raters over time.

Furthermore, our result concerning the human-likeness of g_{kb} clearly demonstrates that semantic knowledge must be brought in to support better, and more human-like, response generation. Large-scale knowledge graphs such as Freebase support many semantic tasks (Jacobs, 1985), and can be used for providing richer context for automatically generating human-like responses.

From a theoretical viewpoint, the system will clearly benefit from rigorous analysis of human interaction in online media. Responses to user-generated content on the Internet share some linguistic characteristics in structure, length and manner of expression. Studying these features theoretically and then examining them empirically using a Turing-like evaluation as presented here can take us a big step in the direction of better generation, and also better understanding of the processes underlying human response generation.

This latter understanding may be complemented with insights into the causes, motivations and intricacies of human interaction in such environments, as studied by sociologists and psychologists. In particular, our preliminary interaction with colleagues from communication studies suggests that the present endeavor nicely complements that of "persuasive computing" (Fogg, 1998; Fogg, 2002), and we hope that this collaboration will lead to valuable synergies.

Finally, bridging the gap between the technical and the theoretical, it would be fascinating to test the responses in the context for which they are generated – social media. Generated texts may be posted as a response to the original article, or shared with a link of the original article, followed by measuring the responses to, and shares of, that response. Such real-world evaluation could indicate that generated responses are indeed believable and engaging, and may better simulate a Turing-like test in which machine-generated responses cannot be distinguished from human responses.

5 Conclusion

We presented a system for generating responses that are directly tied to responders' agendas and document content. To the best of our knowledge, this is the first system to generate subjective responses directly reflecting users' agendas. Our response generation architecture provides an easy-to-use and easy-to-extend solution encompassing a range of NLP and NLG techniques. We evaluated both the human-likeness and the relevance of the generated content, thereby empirically quantifying the efficacy of computer-generated responses compared head-to-head against human responses.

Generating concise, relevant, and opinionated responses that are also human-like is hard — it requires the integration of text-understanding and sentiment analysis, and it is also contingent on the expression of the agents' prior knowledge, reasons and motives. We suggest our architecture and evaluation method as a baseline for future research on generated content that would effectively pass a Turing-like test, and successfully convince humans of the authenticity of generated responses.[5]

Acknowledgments

We thank Yoav Francis for his contribution in the early stages of this research. We further thank our anonymous reviewers for their insightful comments on an earlier draft.

[5] Our code, training data, experimental data (computer and human responses) and analysis scripts are publicly available via `www.tsarfaty.com/nlg-sd/`.

References

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, pages 183–194. ACM.

Teresa M. Amabile. 1981. *Brilliant but Cruel: Perceptions of Negative Evaluators*. Washington, DC: ERIC Clearinghouse.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Douglas M. Bates. 2005. Fitting linear mixed models in R. *R News*, 5:27–30.

Tilman Becker. 2002. Practical, template-based natural language generation with TAG. In *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6)*.

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

David M. Blei. 2012. Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

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.

David DeVault, David Traum, and Ron Artstein. 2008. Practical grammar-based NLG from examples. In *Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 77–85, Stroudsburg, PA, USA. Association for Computational Linguistics.

Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. 2006. An intelligent discussion-bot for answering student queries in threaded discussions. In *Proceedings of Intelligent User Interface (IUI-2006)*, pages 171–177.

B. J. Fogg. 1998. Persuasive computers: Perspectives and research directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '98, pages 225–232, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

B. J. Fogg. 2002. Persuasive technology: Using computers to change what we think and do. *Ubiquity*, December.

Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pages 90–93, 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.

Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. 2013. Predicting and eliciting addressee's emotion in online dialogue. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 964–972, Sofia, Bulgaria, August. Association for Computational Linguistics.

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.*

Paul S Jacobs. 1985. A knowledge-based approach to language production. Technical report, University of California at Berkeley, Berkeley, CA, USA.

Pulkit Kathuria. 2012. Sentiment Classification using WSD, Maximum Entropy and Naive Bayes Classifiers. *https://github.com/kevincobain2000/sentiment_classifier*. Visited March 2014.

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

Marc Langheinrich and Günter Karjoth. 2011. Social networking and the risk to companies and institutions. *Information Security Technical Report. Special Issue: Identity Reconstruction and Theft*, pages 51–56.

Paul Mah. 2012. Tools to automate your customer service response on social media. *http://www.itbusinessedge.com/blogs/smb-tech/tools-to-automate-your-customer-service-response-on-social-media.html*. Visited August 2013.

Chris McConnell. 2012. When brands automate Twitter and Facebook responses I'll revolt. *http://dailytekk.com/2012/06/07/brands-automating-social-media/*. Visited August 2013.

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

Kyoshi Mori, Adam Jatowt, and Mitsuru Ishizuka. 2003. Enhancing conversational flexibility in multimodal interactions with embodied lifelike agent. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, pages 270–272, New York, NY, USA. ACM.

Mickey Nall. 2013. You can't automate social media engagement, argues PRSA's Mickey Nall. *http://www.prmoment.com/1359/you-cant-automate-social-media-engagement-argues-prsas-mickey-nall.aspx*. Visited August 2013.

Karthik Sankaran Narayan, Charles Lee Isbell Jr., and David L. Roberts. 2011. Dextor: Reduced effort authoring for template-based natural language generation. In Vadim Bulitko and Mark O. Riedl, editors, *Proceedings of the Seventh Artificial Intelligence and Interactive Digital Entertainment Conference*. The AAAI Press.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In William W. Cohen and Samuel Gosling, editors, *ICWSM*. The AAAI Press.

Brendan O'Connor, Brandon M. Stewart, and Noah A. Smith. 2013. Learning to extract international relations from political context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1104. The Association for Computer Linguistics.

Jeremiah Owyang. 2012. Brands Start Automating Social Media Responses on Facebook and Twitter. *http://techcrunch.com/2012/06/07/brands-start-automating-social-media-responses-on-facebook-and-twitter/*. Visited August 2013.

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.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

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.

M. Theune, E. Klabbers, J. R. De Pijper, E. Krahmer, and J. Odijk. 2001. From data to speech: A general approach. *Nat. Lang. Eng.*, 7(1):47–86.

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.

Tae Yano, William W. Cohen, and Noah A. Smith. 2009. Predicting response to political blog posts with topic models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 477–485, Stroudsburg, PA, USA. Association for Computational Linguistics.