Chandler: An Explainable Sarcastic Response Generator

Silviu Vlad Oprea[1], **Steven R. Wilson**[1,2], **Walid Magdy**[1]
[1]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
[2]School of Engineering and Computer Science, Oakland University, Rochester, MI, USA
{silviu.oprea,steven.wilson}@ed.ac.uk,wmagdy@inf.ed.ac.uk

Abstract

We introduce Chandler, a system that generates sarcastic responses to a given utterance. Previous sarcasm generators assume the intended meaning that sarcasm conceals is the opposite of the literal meaning. We argue that this traditional theory of sarcasm provides a grounding that is neither necessary, nor sufficient, for sarcasm to occur. Instead, we ground our generation process on a formal theory that specifies conditions that unambiguously differentiate sarcasm from non-sarcasm. Furthermore, Chandler not only generates sarcastic responses, but also *explanations* for why each response is sarcastic. This provides accountability, crucial for avoiding miscommunication between humans and conversational agents, particularly considering that sarcastic communication can be offensive. In human evaluation, Chandler achieves comparable or higher sarcasm scores, compared to state-of-the-art generators, while generating more diverse responses, that are more specific and more coherent to the input.

1 Introduction

The prevalence of sarcasm on the social web (Khodak et al., 2018; Sykora et al., 2020) has motivated more and more computational investigations across the research community. Most focus on textual sarcasm detection (Riloff et al., 2013; Joshi et al., 2016; Wallace et al., 2015; Rajadesingan et al., 2015; Bamman and Smith, 2015; Amir et al., 2016; Hazarika et al., 2018; Oprea and Magdy, 2019): the task of classifying whether or not a given text is sarcastic.

Recently, a new research direction considers sarcasm generation. This is motivated by the potential to create approachable conversational agents. These would be more effective at emulating a human correspondent, considering that sarcasm is a natural part of human discourse (Mishra et al., 2019). The limited amount of work on sarcasm generation is spread across two variants of the task: generating a *sarcastic response* to an input utterance (Joshi et al., 2015); and generating a *sarcastic paraphrase* of an input utterance (Mishra et al., 2019; Chakrabarty et al., 2020).

A major limitation of existing sarcasm generation systems is that they rely on variants of the traditional theory of sarcasm: that the intended meaning concealed by sarcasm is the opposite of the literal meaning. Driven by this assumption, their aim is to generate phrases that either express two incongruous propositions (Joshi et al., 2015; Mishra et al., 2019; Chakrabarty et al., 2020), or express a proposition that is incongruous to the discourse setting (Joshi et al., 2015). However, the traditional theory provides a grounding that is neither necessary, nor sufficient, for sarcasm to occur, as discussed in Section 3. Furthermore, it is less obvious how previous systems that consider the task of generating sarcastic paraphrases, rather than responses, could be used for enabling conversational agents to generate sarcasm.

To overcome these limitations, we first select a formal theory that, from a linguistic-theoretical perspective, specifies devices whose presence is both necessary and sufficient to identify sarcasm, unambiguously differentiating it from non-sarcasm. Grounded on this theory, we propose our sarcasm generation system, *Chandler*[1]. Being grounded in a theory, Chandler is also explainable. That is, we are able to generate not only sarcastic responses, but also *explanations* for why each response is sarcastic. We believe this kind of accountability is crucial for avoiding miscommunication between humans and conversational agents, particularly considering the potentially offensive nature of sarcastic communication (Wilson, 2006).

We employ human annotators on a crowdsourcing platform to evaluate Chandler against state-of-the-art generators, across multiple dimensions.

[1]Inspired by the popular TV sitcom.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 339–349
November 7–11, 2021.

Chandler achieves comparable or higher sarcasm scores, while generating responses that are more diverse, and are perceived as more specific and coherent than those of previous sarcasm generators.

A live demo of the system, allowing users to input an utterance and view sarcastic responses, along with their explanations, is available at `https://bit.ly/ChandlerEMNLP`. We will also release, along with the camera ready version, all inputs, responses, model checkpoints, and the code that implements Chandler, our explainable sarcasm generator, under a Creative Commons CC BY-NC license.

2 Related Work

The earliest work on sarcasm generation is that of Joshi et al. (2015), who introduce SarcasmBot, a sarcastic response generation system. SarcasmBot generates a response based on one of eight possible generators, each containing a set of predefined patterns. The generators do not in fact account for the meaning of the input, rather, they only focus on aspects such as the overall sentiment or presence of swear words.

Mishra et al. (2019) suggest a sarcastic paraphrase generator. They assume that the input is always of negative polarity, and use an unsupervised pipeline of four modules to convert such an input $u^{(-)}$ to a sarcastic version. In the Sentiment Neutralisation module, they filter out negative sentiment words from $u^{(-)}$ to produce $u^{(0)}$. In the Positive Sentiment Induction module, they modify $u^{(0)}$ to convey positive sentiment, producing $u^{(+)}$. Next, in the Negative Situation Retrieval module, they mine a phrase $v^{(-)}$ that expresses a negative situation. $v^{(-)}$ is selected from a set of predefined phrases, based on the similarity to the original input. Finally, the Sarcasm Synthesis module constructs the sarcastic paraphrase from $u^{(+)}$ and $v^{(-)}$.

Chakrabarty et al. (2020) present a similar pipeline. Their R^3 system first employs a Reversal of Valence module, which replaces input words of negative valence with their lexical antonyms using WordNet (Miller, 1995) to produce $u^{(+)}$. Next, it builds an utterance v that is incongruous to $u^{(+)}$, and generates sarcasm from $u^{(+)}$ and v.

Second, they all rely on variants of the traditional theory of sarcasm, which provides a grounding that is neither necessary, nor sufficient, for sarcasm to occur, as discussed in Section 3. Third, the systems of Mishra et al. (2019) and Chakrabarty et al. (2020) are only designed to work with negative inputs. However, sarcastic communication can have many communicative goals, including to praise (Bruntsch and Ruch, 2017), or strengthen friendships (Jorgensen, 1996; Pexman and Zvaigzne, 2004)

3 Linguistic Grounding

The goal of this section is to select a linguistic theory on which to ground the sarcasm generation process.

Gricean Theory In the traditional theory, sarcasm is a form of figurative language that is created by literally saying one thing but meaning the opposite. As Sperber and Wilson (1981) point out, a semantic theory of sarcasm that provides such an explanation would need to provide (a) a mechanism of deriving figurative meaning, (b) and an explanation of why figurative meaning exists in the first place. The traditional theory is incomplete because it does not provide answers to such questions. Moving further Grice (Grice, 1975) sees sarcasm as a blatant flouting of the first maxim of quality ("do not say what you believe is false"), giving rise to a conversational implicature that ensures the cooperative principle is observed. That is, the sarcastic speaker does not figuratively mean, but conversationally implicates the opposite of what they say. A main limitation of the Gricean view is that the flouting is not necessary for sarcasm to occur. For instance, consider sarcastic understatements such as saying "This was not the best movie ever" to mean the movie was bad. Violation is also not sufficient. For instance, it also occurs in the construction of certain stylistic devices, such as metaphors.

Echoic Theories Consider the following scenario:

> [Scenario 1] Alice and Bob go for a walk. Despite Alice's requests, Bob refuses to bring along an umbrella, assuring her it would not rain. Momentarily after leaving the house, it starts raining. Alice to Bob:
>
> (1) a. It's definitely not raining.

Sperber and Wilson (1981) invite us to reconsider the goal of sarcastic utterances. According to the theories discussed so far, Alice's goal is to sarcastically convey her belief about the weather—a belief that is the opposite of what she says. Note, how-

ever, that especially when prosodic or other contextual cues are missing, knowing her belief is a prerequisite for, not a consequence of, recognising her sarcasm. Sperber and Wilson (1981) suggest a different goal that she might have, mainly that of conveying a belief not about the weather, but about the content of 1a itself. In their view, the utterance is not *used*[2] by Alice, but is an echoic *mention* of a previous proposition, mainly the one expressed by Bob's claim that it would not rain. Through the mention, Alice expresses a dissociative attitude towards Bob's claim, perhaps suggesting it was ridiculous of him to expect a dry weather. This *echoic mention theory* offers an explanation for why sarcasm exists in the first place. However, it does not cover all instances of sarcasm. An example of a non-echoic sarcastic utterance is Alice saying:

(1) b. Thanks for leaving the umbrella at home.

The echoic mention theory is also unable to differentiate between sarcastic and non-sarcastic echoic mentions. Several variants of the echoic mention theory have been suggested (Kreuz and Glucksberg, 1989; Wilson and Sperber, 1992; Sperber and Wilson, 1998), however they all suffer from similar limitations. We invite the interested reader to further consult Giora (1995), Kumon-Nakamura et al. (1995), and Utsumi (2000).

Pretense Theories Clark and Gerrig (1984) introduce the *pretense theory* of sarcasm which claims that a sarcastic speaker pretends to be an injudicious person speaking to an imaginary uninitiated audience who would accept the literal interpretation of the speaker's utterance. This way, the speaker expresses a negative attitude towards the pretended person, the imaginary audience, and the situation portrayed through their acting. A variant of the pretense theory is viewing sarcasm as a *joint pretense* of the speaker and the listener (Clark, 1996). That is, instead of the speaker pretending to be an imaginary person, both interlocutors pretend to be in an imaginary situation in which they are performing a serious communicative act directed at the listener. Sarcasm is caused by the joint pretense of the speaker and the listener that the imaginary situation is taking place. Both theories fail to distinguish sarcasm from non-sarcastic pretense, such as parody. Their assumptions are, therefore, not sufficient to explain sarcasm. They are also not necessary. An argument in this direction is provided by Utsumi (2000, p. 1782).

[2]See Sperber and Wilson (1981) for a discussion on the use–mention distinction.

3.1 Implicit Display Theory

The theories reviewed so far make assumptions about the nature of sarcasm that are neither necessary, nor sufficient, for sarcasm to occur. We now introduce the Implicit Display Theory (IDT) (Utsumi, 1996), which focuses specifically on making the distinction between sarcasm and non-sarcasm. Because of this, we chose it to serve as a grounding for our generation process. We provide a brief introduction here, but invite the interested reader to consult (Utsumi, 2000) for an overview of how it overcomes the limitations of previous theories.

The IDT first defines the concept of an ironic environment. We say a situation in which an utterance occurs is surrounded by an ironic environment if the discourse context includes the following components:

1. The speaker has expectation Q at time t_0;
2. Q fails at time $t_1 > t_0$;
3. The speaker has a negative attitude towards the failure of Q.

In Utsumi (1996)'s view, such a situation within the discourse context facilitates the use of sarcasm. Note that the negative attitude could have several intensities, could be serious, or joking. Note also that the idea of linking sarcasm to an expectation is not new to Utsumi (1996), rather it is supported by previous work (Kreuz and Glucksberg, 1989; Kumon-Nakamura et al., 1995).

From here, according to the IDT, an utterance is sarcastic if and only if it is given in a situation surrounded by an ironic environment and it implicitly displays all three components of the ironic environment. Implicit display is realised if the utterance:

1. alludes to the speaker's failed expectation Q;
2. includes pragmatic insincerity, by intentionally violating one of the pragmatic principles;
3. implies (indirectly expresses) the speaker's negative attitude towards the failure of Q.

The pragmatic principles that we are referring to include, among others, Grice's maxims (Grice, 1975), and the felicity conditions for well-formed speech acts (Searle and Searle, 1969).

A final claim of the theory is that sarcasm is a prototype-based category characterised by implicit display. That is, the degree of sarcasm of an utter-

ance is proportional to how many implicit display conditions the utterance meets.

4 Methodology

The IDT directly suggests an algorithm for sarcasm generation that identifies an ironic environment, then creates an utterance that implicitly displays it. We now discuss how we implement each step.

Ironic Environment Let U_{in} be an input text to our system. Herein, we assume the expectation Q that is part of the ironic environment negates what U_{in} proposes. For instance, say U_{in} expresses the event $P = [$<user> wins the marathon$]$. We assume $Q = \neg P = [$<user> does not win the marathon$]$. As we shall see, the algorithm we suggest will not, in fact, require us to formulate Q, but it relies on the above assumption.

Allusion to Q Following Utsumi (2000), we define allusion in terms of coherence relations, similar to the relations of rhetorical structure theory (RST) (Mann and Thompson, 1987). That is, if U is an utterance that expresses proposition α, we say U alludes to the expectation Q if and only if there is a chain of coherence relations from α to Q. So, we need to first select a proposition α to either start or end the coherence chain, then specify the chain between α and Q, and formulate U such that it expresses α. We suggest defining such α as objects of if-then relations, where the subject is P, the proposition expressed by input text U_{in}. That is, relations of the form "if P then α" should hold. To infer α given U_{in}, we use COMET (Bosselut et al., 2019), an adaptation framework for constructing commonsense knowledge. Specifically, we use the COMET variant fine-tuned on ATOMIC (Sap et al., 2019), a dataset of typed if-then relations. [3]. COMET inputs the subject of the relation, along with the relation type, and outputs the relation object. In our case, the subject is U_{in}, and we set α to the output.

We leverage four relation types. In the examples that follow, assume the input text is $U_{\text{in}} =$'<user> won the marathon': (1) **xNeed**: the object α of a relation of this type specifies an action that the user needed to perform before the event took place, e.g. "if U_{in} then $\alpha = [$*xNeed* to train hard$]$"; (2) **xAttr**: the object α specifies how a user that would perform such an action is seen, e.g. "if P then $\alpha = [$*xAttr* competitive$]$"; (3) **xReact**: the object α specifies how the user could feel as a result of the event, e.g. "if P then $\alpha = [$*xReact* happy$]$"; and (4) **xEffect**: the object specifies a possible effect that the action has on the user, e.g. "if P then $\alpha = [$*xEffect* gets congratulated$]$". In Table 1 we show, for each relation type, the coherence chains between the relation object α and the failed expectation Q. Under these conditions, to generate an utterance U that alludes to Q, we simply need to choose U to expresses α.

[3] We use the COMET checkpoint published at http://bit.ly/comet-checkpoints.

Algorithm 1: Generate sarcastic response

input: utterance U_{in};
ironic environment
  Let $Q := \neg P$ be the failed expectation;
implicit display
  Choose an if-then relation type τ from *xNeed*, *xAttr*, *xReact*, and *xEffect*;
  Let $\alpha = \text{COMET}(U_{\text{in}}, \tau)$;
return response U that expresses *emotion*$(\neg\alpha)$;

Pragmatic insincerity The second requirement for implicit display is that the utterance generated U should include pragmatic insincerity. In this paper, we focus on violating Grice's maxim of quality (Grice, 1975), where we aim for the propositional contents of U (generated utterance) and U_{in} (input text) to be incongruous. To achieve this, we first choose an if-then relation type, then infer the relation object α from U_{in} using COMET, and construct U to express $\neg\alpha$. For instance, if $U_{\text{in}} =$'<user> won the marathon', and we have chosen the *xAttr* relation type, U could be chose to express $\neg\alpha = [$<user> is not competitive$]$.

Negative attitude To fulfill the last requirement of implicit display, the utterance generated should imply a negative attitude towards the failure of the expectation Q. As pointed out by Utsumi (1996), this can be achieved by embedding verbal cues usually associated with such attitudes, including hyperbole and interjections.

Logical form and explainability At this point we formulate Algorithm 1 for generating a sarcastic response U, given an input utterance U_{in} that expresses proposition P. We refer to *emotion*$(\neg\alpha)$ as the *logical form* of the sarcastic response we generate. Here, *emotion* is a function that augments $\neg\alpha$ to express a negative attitude. Note that the logical form, together with the coherence chain between α and the failed expectation Q, provide a complete explanation for *how* and *why* sarcasm occurs. The explanation is $\epsilon = (\text{emotion}(\neg\alpha), \mathcal{C})$, where is the coherence chain from α to Q. The co-

relation type	example relation	coherence chain
xNeed	if P then α = [xNeed to train hard]	volitional-cause(α, P) and contrast(P, Q)
xAttr	if P then α = [xAttr competitive]	condition(α, I_P) $\wedge$ purpose(I_P, P) $\wedge$ contrast(P, Q)
xReact	if P then α = [xReact happy]	contrast(Q, P) $\wedge$ volitional-result(P, α)
xEffect	if P then α = [xEffect gets congratulated]	contrast(Q, P) $\wedge$ non-volitional-result(P, α)

Table 1: Coherence chains between the object α of an if-then relation and the failed expectation Q, for each relation type, as discussed in Section 4. Here, P is the proposition expressed by the input text U_{in}. In the examples, U_{in} ='<user> won the marathon'.

herence chain for each relation type can be selected from Table 1. This makes our sarcasm generation process accountable.

Logical Form to Text To convert the logical form to text, we rely on predefined patterns for each if-then relation type. As a running example, assume the input utterance U_{in} ='<user> won the marathon' and the chosen relation type is *xAttr*. Say α = COMET(U_{in}, *xAttr*) = [*xAttr* competitive]. The logical form is *emotion*($\neg$[*xAttr* competitive]). We construct an intermediate utterance U^0_{out} using the rule <*user*> <*verb*> *competitive*, where <*verb*> is a verb specific to each relation type. In our example, U^0_{out} could be '<user> is competitive'. From U^0_{out}, we generate a sarcastic response U_{out} to U_{in} as follows. We first apply a rule-based algorithm to generate the negation of U^0_{out} in a manner similar to Chakrabarty et al. (2020), discussed in Section 2. The result could be '<user> is not competitive', expressing $\neg$[*xAttr* competitive]. Next, in a pattern-based manner, we augment this with hyperbole and interjections, as indicated by Utsumi (2000), to get U_{out}, expressing *emotion*($\neg$[*xAttr* competitive]). This could be '<user> is definitely not competitive, yay!'. A full list of patterns is shown in the Appendix A.

In the running example we focused on the *xAttr* relation type. Recall there are four relation types that we consider, *xNeed*, *xAttr*, *xReact*, and *xEffect*. As such, for each input text U_{in}, we generate 4 responses, one for each relation type. We use the pattern Ch-<relation > to refer to each response of our system, *Chandler*. For instance, Ch-xAttr refers to U_{out} built considering the *xAttr* relation, while Ch-xNeed refers to U_{out} built considering the *xNeed* relation.

Note that other strategies for converting the logical form of sarcasm to text are possible. For instance, using policy-based generation with external rewards (Mishra et al., 2019) might have lead to higher perceived sarcasticness of our generated responses. We leave this to future work. Our goal was to provide a theory-based, explainable, generation framework.

5 Experiments

5.1 Setup

To evaluate Chandler, we built a survey[4] that we published on the Prolific Academic[5] crowdsourcing platform. In the survey, we presented annotators with the input text U_{in}, along with the responses produced by Chandler-xNeed, Chandler-xAttr, Chandler-xReact, and Chandler-xEffect.

We also included a response from DialoGPT (Zhang et al., 2020), a recent dialogue system that is not built to be sarcastic; a response produced by SarcasmBot, the sarcastic response generator of Joshi et al. (2015); and a response produced by R^3, the state-of-the-art sarcastic paraphrase generator of Chakrabarty et al. (2020). While not designed to produce responses, we applied R^3 to the output of DialoGPT to get a sarcastic rephrase of a response to the input.

As inputs, we selected texts from the corpus published by Wilson and Mihalcea (2019). The corpus contains short texts (extracted from tweets) where users describe actions they performed. We compute the sentiment polarity of each text using the classifier of Barbieri et al. (2020), a RoBERTa model (Liu et al., 2019) fine-tuned on the tweet sentiment dataset of Rosenthal et al. (2017). Next, we form five partitions of 50 texts each: *very negative* and *very positive*, containing the top 50 texts based on their negative and positive probabilities, respectively; *negative*, containing random texts for which the probability of being negative was higher that the probabilities of being positive or neutral; and *positive* and *neutral*, partitions that we formed analogously to how we formed the *negative* partition. Our final input dataset contains 250 texts. For each input, we collected 3 annotations for its

[4]Participant information sheet is shown in Appendix C.
[5]https://prolific.co

system	response
DialoGPT	I'm not sure if you're being sarcastic or not.
DialoGPT+R^3	I'm sure if you're being sarcastic or not. No one has yet been hurt.
SarcasmBot	That is a very useful piece of information! LMAO
Ch-xNeed	Yay! Good job not knowing how to write.
Ch-xAttr	Yay! You're not a very unintelligent person, that's for sure.
Ch-xReact	You're not feeling very embarrassed right now, that's for sure. Yay!
Ch-xEffect	You're not really going to sigh in frustration right now, that's for sure. Brilliant!

Table 2: Responses generated by all systems to the utterance "I ran out of characters :drooling_face:", as discussed in Section 5.1.

System	sarc.	hum.	coh.	spec.	diversity
DialoGPT (non-sarcastic)	0.6	0.3	2.3	2.0	0.92
DialoGPT+R^3	0.8	0.3	0.9	1.3	0.92
SarcasmBot	2.5	0.8	1.4	0.9	0.14
Ch-xNeed	1.9	0.6	1.3	1.6	0.80
Ch-xAttr	2.1	0.6	1.3	1.4	0.80
Ch-xReact	1.7	0.4	1.0	1.0	0.35
Ch-xEffect	1.6	0.5	1.1	1.3	0.67

Table 3: Means of the sarcasm, humour, specificity, and coherence scores provided by annotators, for each variant of Chandler (Ch), as discussed in Section 5.2. Diversity is the ratio of unique responses generated for our 250 inputs.

responses.

Table 2 shows an example input utterance, along with responses from all systems.

All in all, each survey instance contained a specific input text, and seven responses generated as mentioned above and presented in a random order. In the survey, we asked annotators to evaluate each response across four dimensions: (1) Sarcasm: How sarcastic is the response? (2) Humour: How funny is the remark? (3) Coherence: How coherent is the remark to the input? It is coherent if it sounds like sensible response that a person might give in a real conversation; and (4) Specificity: How specific is the remark to the input? It is not specific if it can be used as a response to many other inputs. Each dimension ranged from 0 to 4, in line with previous work (Chakrabarty et al., 2020).

5.2 Results

In Table 3 we show mean sarcasm, humour, specificity, and coherence scores provided by annotators for each variant of Chandler, across all inputs.

We have four strategies for alluding to the failed expectation, depending on the relation type considered. We notice the highest sarcasm score is achieved by Ch-xAttr, followed by Ch-xNeed, Ch-xReact and Ch-xEffect. Out of the allusion strategies selected, the responses perceived as most sarcastic are those that mention attributes of the user. Similarly, we notice that, among variants of Chandler, those that use the xAttr and xNeed relations are perceived and the most coherent and specific to the input, and achieve the highest humour score.

Chandler achieves lower specificity and coherence scores compared to DialoGPT, which is to be expected considering that DialoGPT is not designed to conceal the intended meaning using sarcasm. The sarcasm score, however, for all variants of Chandler, is considerably higher compared to DialoGPT. The situation is similar when comparing Chandler to DialoGPT+R^3.

When comparing to SarcasmBot, while specificity is considerably higher for most variants of Chandler, and coherence is similar, sarcasm score is slightly lower. In particular, the most sarcastic variant of Chandler, Ch-xAttr, achieves a sarcasm score of 2.1, compared 2.5 achieved by SarcasmBot. This is expected, considering that SarcasmBot provides responses from a fixed set of responses that were carefully curated for sarcasm. However, using SarcasmBot in the real world is not practical, as the original authors point out (Joshi et al., 2015). When analysing its outputs, we noticed a very low diversity, as shown in Table 3, where we define diversity as ratio of unique responses generated across our 250 inputs. In particular, SarcasmBot produced a total of only 28 unique responses[6]. In a real scenario of a user interacting with a conversational agent, the user might not appreciate repeatedly receiving the same response, that is not even specific to the meaning of the input. Indeed, in our experiments, we noticed that most of the time a fallback generator of SarcasmBot was employed, returning the simple concatenation of a random positive phrase to a random negative one, from a set of predefined phrases that have no specific connection to the input.

6 Conclusion

We have presented Chandler, a linguistically informed framework for generating sarcastic responses to an input utterance. Chandler is the first such system that does not rely exclusively on predefined patterns, and focuses on explainable generation, grounded on a linguistic theory of sarcasm that overcomes the limitations of previous theories assumed by previous sarcasm generators.

[6]All 28 responses are listed in the Appendix B.

Acknowledgments

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1); the University of Edinburgh; and The Financial Times.

References

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mario J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *CoNLL*, pages 167–177. ACL.

David Bamman and Noah A. Smith. 2015. Contextualized sarcasm detection on twitter. In *ICWSM*, pages 574–577. AAAI Press.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. ACL.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *ACL*, pages 4762–4779. ACL.

Richard Bruntsch and Willibald Ruch. 2017. Studying irony detection beyond ironic criticism: Let's include ironic praise. *Frontiers in Psychology*, 8:606.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. Rˆ3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *ACL*, pages 7976–7986. ACL.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Herbert H. Clark and Richard J. Gerrig. 1984. On the pretense theory of irony. *Journal of Experimental Psychology: General*, 113(1):121–126.

Rachel Giora. 1995. On irony and negation. *Discourse Processes*, 19(2):239–264.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, Cambridge, UK.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. In *COLING*, pages 1837–1848. ACL.

Julia Jorgensen. 1996. The functions of sarcastic irony in speech. *Journal of Pragmatics*, 26(5):613–634.

Aditya Joshi, Anoop Kunchukuttan, Mark James Carman, and Pushpak Bhattacharyya. 2015. Sarcasmbot: An open-source sarcasm-generation module for chatbots. In *WISDOM at KDD*. ACM.

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are word embedding-based features useful for sarcasm detection? In *EMNLP*, pages 1006–1011. ACL.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *LREC*. ELRA.

Roger J. Kreuz and Sam Glucksberg. 1989. How to Be Sarcastic: The Echoic Reminder Theory of Verbal Irony. *Journal of Experimental Psychology: General*, 118(4):374–386.

Sachi Kumon-Nakamura, Sam Glucksberg, and Mary Brown. 1995. How about another piece of pie: The allusional pretense theory of discourse irony.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

William C. Mann and Sandra A. Thompson. 1987. *Rhetorical Structure Theory: Description and Construction of Text Structures*, pages 85–95. Springer Netherlands, Dordrecht.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. 2019. A modular architecture for unsupervised sarcasm generation. In *EMNLP-IJCNLP*, pages 6144–6154. ACL.

Silviu Oprea and Walid Magdy. 2019. Exploring author context for detecting intended vs perceived sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859, Florence, Italy. Association for Computational Linguistics.

Penny M. Pexman and Meghan T. Zvaigzne. 2004. Does irony go better with friends? *Metaphor and Symbol*, 19(2):143–163.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *WSDM*, pages 97–106. ACM.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, pages 704–714. ACL.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. ACL.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

John R Searle and John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge University Press, Cambridge, UK.

Dan Sperber and Deirdre Wilson. 1981. Irony and the use-mention distinction. *Philosophy*, 3:143–184.

Dan Sperber and Deirdre Wilson. 1998. Irony and relevance: A reply to seto, hamamoto and yamanashi. In R. Carston and S. Uchida, editors, *Relevance theory: Applications and implications*, pages 289–293. Benjamins, Amsterdam.

Martin Sykora, Suzanne Elayan, and Thomas W Jackson. 2020. A qualitative analysis of sarcasm, irony and related #hashtags on twitter. *Big Data & Society*, 7(2):2053951720972735.

Akira Utsumi. 1996. Implicit display theory of verbal irony: Towards a computational model of irony. In *Proceedings of the International Workshop on Computational Humor (IWCH'96)*.

Akira Utsumi. 2000. Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, 32(12):1777–1806.

Byron C. Wallace, Do Kook Choe, and Eugene Charniak. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *ACL*, pages 1035–1044. ACL.

Deirdre Wilson. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722–1743.

Deirdre Wilson and Dan Sperber. 1992. On verbal irony. *Lingua*, 87(1):53–76.

Steven Wilson and Rada Mihalcea. 2019. Predicting human activities from user-generated content. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2572–2582, Florence, Italy. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278. ACL.

A Logical Form to Text Patterns

In this Section we show the patterns used by Chandler to convert the logical form of sarcasm to text, as discussed in Section 4. We show patterns for each if-then relation type, *xNeed*, *xAttr*, *xReact*, and *xEffect*.

In the patterns below, <inten> is an intensifier, <suff_inten> is an intensifier added at the end of a phrase, <pos> is a positive emotion word, and <interj> an interjection. Inspired by (Utsumi, 2000) and (Joshi et al., 2015), each of these were randomly chosen from the following sets:

- <inten> : [very]
- <suff_inten> : [for sure]
- <pos> : [Good job, Well done]
- <intrj> : [Yay!, Brilliant!]

<obt> below is the object of the corresponding if-then relation object, as provided by COMET when taking in the input tweet.

A.1 Patterns for the Complete Version of Chandler

xNeed patterns:

- You didn't <obt> , that's <suff_inten> . <pos> !

xAttr patterns:

- <interj> You're not <inten> <obt> , that's <suff_inten> .
- <interj> <pos> not being <obt> .
- <interj> You're not a very <obt> person that's <suff_inten> ."

xReact patterns:

- You're not feeling <inten> <obt> right now, that's <suff_inten> . <interj>

xEffect patterns:

- You're not <inten> going to obt_inf right now, that's <suff_inten> . <interj>

B SarcasmBot Responses

As discussed in Section 5.2, we noticed SarcasmBot produced a total of only 28 unique responses to our set of 250 inputs. Here they are:

1. Unbelievable that you just said 'sucky'! You are really very classy!
2. Awesome!
3. Brilliant!
4. Let's party!
5. Oh you poor thing!
6. You owe me a drink for that awesome piece of news!

7. Wow, you said 'sucks', didn't you? Your mom will be really proud of you!
8. Wow, you said 'suck', didn't you? Your mom will be really proud of you!
9. I'd feel terrible if I were you!
10. You are such a simple person!
11. Aww!! That's so adorable!
12. That deserves an applause.
13. I am so sorry for you!
14. Yay! Yawn!
15. How exciting! Yawn!
16. How exciting! *rolls eyes*
17. Wow! *rolls eyes*
18. Yay! *rolls eyes*
19. Yay! LMAO
20. Wow! Yawn!
21. How exciting! LMAO
22. Wow! LMAO
23. That is a very useful piece of information! *rolls eyes*
24. That is a very useful piece of information! LMAO
25. That is a very useful piece of information! Yawn!
26. Unbelievable that you just said 'sobbing'! You are really very classy!
27. Unbelievable that you just said 'sucks'! You are really very classy!
28. Unbelievable that you just said 'bloody'! You are really very classy!

C Participant Information Sheet

C.1 What will I do?

Imagine someone (we'll call them PersonX), makes a statement. You will be shown a few responses to that statement. The responses were generated by chatbots (computer programs). Some sentences talk about sensitive topics, such as tragic life events. Responses to such sentences could be potentially inappropriate, or even offensive or harmful. Unfortunately, chatbots do not understand whether or not a topic is sensitive for a human. Please be fully aware of this when accepting to take part in our study.

For each response, you will be asked:

1. How sarcastic you find the response? (0 - not sarcastic, 3 - very sarcastic)

2. How funny you find the response? (0 - not funny, 3 - very funny)

3. How specific is the response to PersonX's statement? The response is specific if it mentions details that show a good understanding of PersonX's statement and its implications. Otherwise it's general. (0 - very general, 3 - very specific).

4. How coherent is the response to PersonX's statement? The response is coherent if it makes sense as a response. That is, it's a clear and sensible response that someone might actually give. It does not matter if it's specific or general. (0 - not coherent, 3 - very coherent).

Let's take a quick example. In this example, imagine that PersonX's statement is "I went to the grocery store". Here are some responses about this statement.

About being specific:

- "That's great." - Very general response. You can say this as a response to pretty much anything.

- "Nice to hear you are enjoying this sunny day." - General response. It does provides some details about the day (that it's sunny). However, those details are not uniquely related to PersonX's statement.

- "You must be tired." - More specific response. It shows an understanding that going somewhere (anywhere at all) may cause tiredness.

- "You probably bought a lot of vegetables." - Specific response. It shows an understanding of what a grocery store is. That is, a place where you can probably buy vegetables.

- "You must have been quite hungry for carrots." - Very specific response. It shows an understanding of what a grocery store is, about what carrots are, and about the link between carrots and the store (mainly, that carrots are sold there).

About being coherent:

- "I'm cold." - Not coherent. It has nothing to do with PersonX's statement

- "I went to the grocery store". It's not a suitable response that someone would normally give.

- "I had such a wonderful dream last night, there were a lot of awesome cars painted blue." - Not coherent. It does not make sense as a response to PersonX's statement.

- "I sometimes dream about eating carrots." - More coherent response. Someone might sometimes say this as a response, although it's not a common response.

- "OK thanks." - Very coherent. One might actually say this as a response. Notice it's not specific to PersonX's statement. You can say it as a response to many other statements. Still, it's coherent to PersonX's statement. Thanks a lot for getting me those carrots, I'll pay you back next week. - Very coherent and very specific to PersonX's statement.

C.2 Participant Information Sheet and Consent Form

- Principal investigator: Prof. Walid Magdy

- Researcher collecting data: Silviu Oprea

- Funder (if applicable): EPSRC, Financial Times

This study is in the process of being certified according to the Informatics Research Ethics Process, RT number 2019/87618. Please take time to read the following information carefully. You should keep this page for your records.

C.3 Who are the researchers?

We are the Social Media Analysis and Support for Humanity (SMASH) group, a research group that brings together a range of researchers from the University of Edinburgh in order to build on our existing strengths in social media research. This research group focuses on mining structures and behaviours in social networks. The principal investigator is Prof. Walid Magdy.

C.4 What is the purpose of the study?

This study aims to understand what linguistic style people associate with sarcasm.

C.5 Why have I been asked to take part?

We target everyone registered as living in the United Kingdom on the Prolific Academic platform.

C.6 Do I have to take part?

No—participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. Your rights will not be affected. If you wish to withdraw, contact the PI. We will stop using your data in any publications or presentations submitted after you have withdrawn consent. However, we will keep copies of your original consent, and of your withdrawal request.

C.7 What will happen if I decide to take part?

You will be asked to fill in a survey. The flow of the survey is the following:

- You will be shown a short text (originating from a tweet) and asked whether it is, in your view, appropriate to respond sarcastically to that text.

- If you say "no", you will be shown another text. The process will repeat until you say "yes" or 10 texts have been shown.

- If you say "yes":
  - You will be shown 7 responses to the text that you selected;
  - For each response, you will be asked to specify, on a scale from 1 to 5: (a) How sarcastic it is; (b) How funny it is; (c) How coherent it is to the original text; It is coherent if it sounds like a reasonable response that a person might give. (d) How specific it is to the original text; It is specific if it mentions details about the original text, or its implications, that make this response not appropriate as a response to many other texts.

We estimate it will take around 3 minutes to complete the survey.

C.8 Compensation

You will be paid £0.38 for your participation in this study.

C.9 Are there any risks associated with taking part?

Please note: some of the texts that you will see include content that you might consider sensitive, or might trigger unwanted memories. For instance,

they might mention losing a family member, losing friends, break-ups, failure in exams, or health issues.

C.10 Are there any benefits associated with taking part?

Financial compensation of £0.38.

C.11 What will happen to the results of this study?

The results of this study may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a minimum of 2 years.

C.12 Data protection and confidentiality

Your data will be processed in accordance with Data Protection Law. Throughout your entire interaction with us, the only information collected about you specifically is your Prolific Academic identification number. This data will only be viewed by the team members of the SMASH group, listed here: http://smash.inf.ed.ac.uk. All other data, including the responses you provide, and the amount of time you took to fill in the survey, will be made public on the internet as part of Open Science, available to be indexed by search engines. The Open Science initiative is described here: https://en.wikipedia.org/wiki/Open_science.

C.13 What are my data protection rights?

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. However, we will have no control for the data that will be made public, as specific in the previous section. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk. For general information about how we use your data, go to: edin.ac/privacy-research.

C.14 Who can I contact?

If you have any further questions about the study, please contact the lead researcher, Silviu Oprea, silviu.oprea@ed.ac.uk. If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

C.15 Updated information

If the research project changes in any way, an updated Participant Information Sheet will be made available on http://web.inf.ed.ac.uk/infweb/research/study-updates.

C.16 Consent

By proceeding with the study, you agree to all of the following statements:

- I have read and understood the above information.
- I understand that my participation is voluntary, and I can withdraw at any time.
- I consent to my anonymised data being used in academic publications and presentations, as well as published publicly on the internet, as part of Open Science.
- I am aware that I will see potentially offensive, harmful, or hurtful content.
- I allow my data to be used in future ethically approved research.