Harvesting Creative Templates for Generating Stylistically Varied Restaurant Reviews

Shereen Oraby, Sheideh Homayon, and **Marilyn Walker**
Natural Language and Dialogue Systems Lab
University of California, Santa Cruz
{soraby,shomayon,mawalker}@ucsc.edu

Abstract

Many of the creative and figurative elements that make language exciting are lost in translation in current natural language generation engines. In this paper, we explore a method to harvest templates from positive and negative reviews in the restaurant domain, with the goal of vastly expanding the types of stylistic variation available to the natural language generator. We learn hyperbolic adjective patterns that are representative of the strongly-valenced expressive language commonly used in either positive or negative reviews. We then identify and delexicalize entities, and use heuristics to extract generation templates from review sentences. We evaluate the learned templates against more traditional review templates, using subjective measures of *convincingness*, *interestingness*, and *naturalness*. Our results show that the learned templates score highly on these measures. Finally, we analyze the linguistic categories that characterize the learned positive and negative templates. We plan to use the learned templates to improve the conversational style of dialogue systems in the restaurant domain.

#	Stars	Review
1	1/5	This place is probably the worst thing that ever happened to the history of the known world. [...] The food, however, I initially would want to call unremarkable but I can't. I can't call it unremarkable because it is so incredibly remarkably terrible. [...]
2	2/5	Can't say anything about the food, as we were never served. We never saw a server, even after sitting at our table for 15 minutes. Unacceptable.
3	3/5	I was back here a couple of days ago with my family. And although I remember The food being a lot better than this time around. I was kind of disappointed. The service was okay since I had no Jose this time. Nothing to mention here just refills chips salsa and beverages when you need and food when it's ready.
4	4/5	I would eat here everyday if I didn't think I'd end up 400 pounds... Minus 1 star because each time I've been here the service has kinda sucked and orders have been messed up. Regardless, their fried chicken on waffles topped with syrup and a slice of Red Velvet cake to top it off......... is sooooooo heavenly.
5	5/5	I only have one warning about this restaurant. The food is so amazing that you cannot eat Mexican food anywhere else. [...] I had chicken and beef enchiladas which had homemade corn tortillas and the most tender meat I had ever tasted. [...] I will be a customer for life here!

Table 1: Restaurant Reviews by Rating from the Yelp Dataset Challenge Corpus

1 Introduction

The restaurant domain has been one of the most common applications for spoken dialogue systems for at least 25 years (Polifroni et al., 1992; Whittaker et al., 2002; Stent et al., 2004; Devillers et al., 2004; Gasic et al., 2008). There has been a tremendous amount of previous work on natural language generation of recommendations and descriptions for restaurants (Howcroft et al., 2013; Wen et al., 2015; Novikova et al., 2016), some of which has even focused on generating stylistically varied restaurant recommendations (Higashinaka et al., 2007b; Mairesse and Walker, 2010; Dethlefs et al., 2014). Given this, it is surprising that previous work has not especially noted that restaurant reviews are a fertile source of creative and figurative language. For example, consider the elaborate descriptions in the restaurant reviews in Table 1[1], e.g. phrases such as *worst thing that*

[1] Reviews from the Yelp 2016 dataset challenge: https://www.yelp.com/dataset_challenge

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 28–36
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

ever happened in the history of the known world along with *incredibly remarkably terrible* (Row 1), *eat here everyday if I didn't think I'd end up 400 pounds* and *sooooooo heavenly* (Row 4), and *food so amazing you cannot eat [...] anywhere else* (Row 5). These phrases express extremely valenced reactions to restaurants, their menu items, and related attributes, using figurative language.

The creativity exhibited in these user-generated restaurant reviews can be contrasted with natural language generation (NLG) for the restaurant domain. Methods for NLG typically begin with a structured meaning representation (MR), as shown in Table 2, and map these meaning representations into surface language forms, using a range of different methods, including template-based generation, statistically trained linguistically-informed NLG engines, and neural approaches (Bangalore and Rambow, 2000; Walker and Rambow, 2002). These approaches vary in the degree to which they can generate syntactically and semantically correct utterances, but in most cases the stylistic variation they can generate is extremely limited. Table 2 illustrates sample restaurant domain utterances produced by recent statistical/neural natural language generators (Higashinaka et al., 2007a; Mairesse and Walker, 2007; Wen et al., 2015; Novikova et al., 2016; Dusek and Jurcícek, 2016).

One of the most prominent characteristics of restaurant reviews in the Yelp corpus is the prevalent use of hyperbolic language, such as the phrase *"incredibly remarkably terrible"* in Table 1. Hyperbole is often found in persuasive language, and is classified as a form of figurative language (McCarthy and Carter, 2004; Cano Mora, 2009). Colston and O'Brien describe how an event or situation evokes a scale, and how hyperbole exaggerates a literal situation, introducing a discrepancy between the "truth" and what is said (Colston and Keller, 1998; Colston and O'Brien, 2000). Hyperbole moves the strength of a statement up and down the scale, away from the literal meaning, where the degree of movement reflects the degree of contrast or exaggeration. Depending on what they modify, adverbial intensifiers like *totally, absolutely,* and *incredibly* can shift the strength of the assertion to extreme negative or positive.

Similarly, Kreuz and Roberts (1995) describe a standard frame for hyperbole in English where an adverb modifies an extreme, positive adjective, e.g. *"That was* **absolutely amazing***!"* or *"That was* **simply the most incredible** *dining experience in my entire life."* Such frames can be seen in the reviews in Table 1, but we also see many other idiomatic hyperbolic expressions such as *out of this world* (Cano Mora, 2009).

Our goal is to develop a natural language generator for the restaurant domain that can harvest and make use of these types of stylistic variations. We explore a data-driven approach to automatically select stylistically varied utterances in the restaurant review domain as candidates for review construction. We empirically learn hyperbolic adjective patterns that are highly correlated with two classes (positive and negative reviews). Using different resources, we also identify and delexicalize *restaurant, cuisine, food, service,* and *staff* entities, and select short, single-entity utterances that are simple to templatize.

Our overall approach is thus similar to Higashinaka et al. (2007a,b), who describe a method for harvesting an NLG dictionary from restaurant reviews, however our focus on learning expressive language, in particular hyperbole as a type of figurative language, is novel. Our framework consists of the following steps:

1. Collect a large number of strongly positive and strongly negative reviews in the restaurant domain;
2. Use a linguistic pattern learner to identify linguistic frames that use hyperbole;
3. Create generation templates from the identified linguistic patterns and infer their contexts of use;
4. Learn to rank the generation templates for convincingness and quality.

We see Steps 1 to 3 as the overgeneration phase, aimed at vastly expanding the types of stylistic variation possible, while Step 4 is the ranking phase, in a classic overgenerate and rank NLG architecture (Langkilde and Knight, 1998; Rambow et al., 2001). We focus in this paper on Steps 1 to 3, expecting to improve these steps before we move on to Step 4.

Thus, in this paper, we conducted an evaluation experiment to compare three different types of NLG templates: pre-defined BASIC templates similar to those used in current NLG engines for the restaurant domain (Walker et al., 2007; Wen et al., 2015), the basic templates stylized with

MR	NLG	System
name[Emilios] foodquality[excellent] location[city centre] cuisine[Italian] decor[decent] service[decent]	*Emilios decor and service are both decent, but its food quality is nothing short of excellent. It serves Italian food and its in the City Centre.*	Seq2Seq NLG (Nayak et al., 2017)
name[The Eagle] eattype[coffee shop] pricerange[moderate] customerrating[3/5] kidsFriendly[yes] near[Burger King]	*The three star coffee shop, The Eagle, gives families a mid-priced dining experience featuring a variety of wines and cheeses. Find The Eagle near Burger King.*	Human annotators (Novikova et al., 2016)
name[X] type[placetoeat] near[Y] food[Continental] food[French]	*X is a restaurant providing french and continental food, near Y.*	Seq2Seq Generation (Dusek and Jurcícek, 2016)
name[Red door cafe] goodformeal[breakfast] area[cathedral hill] kidsallowed[no]	*Red door cafe is a good restaurant for breakfast in the area of cathedral hill and does not allow children .*	LSTM Generation (Wen et al., 2015)
name[Le Marais] cuisine[French] foodquality[good] service[good] decor[decent] price[$44]	*I am sure you would like Le Marais, you know. The atmosphere is acceptable, the servers are nice and it's a fresh, kosher and steak house place. Actually, the food is good, even if its price is 44 dollars.*	PERSONAGE (Mairesse and Walker, 2007)
name[Babbo] foodquality[superb] service[excellent] decor[superb]	*The food is phenomenal and the atmosphere is very unique. Babbo has excellent service. It has the best overall quality among the selected restaurants.*	Unsupervised Method for Lexicon Learning (Higashinaka et al., 2007a)

Table 2: Example Meaning Representations (MR) and Corresponding Natural Language Generation (NLG) Output in the Restaurant Domain

our learned patterns for more HYPERBOLIC templates, and finally a class of CREATIVE templates that incorporate full sentence templates from user reviews. Our expectation was that many of the CREATIVE templates would fail to be appropriate to their contexts, but that our HYPERBOLIC templates would be both appropriate and more interesting and convincing than the BASIC templates. However, our results show that our creative templates are preferred as more convincing, interesting, and natural across the board. We discuss how we can use quantitative metrics associated with the learned templates for future ranking, and analyze characteristic linguistic categories in each class.

2 Data

Our restaurant review data comes from the Yelp dataset challenge, which includes 144K businesses with over 4.1M reviews. We randomly select 10K businesses located in the US that are classified as restaurants, resulting in a set of around 40K reviews. The data consists of around 4K 1 stars, 3.8K 2 stars, 5.6K 3 stars, 11.3K 4 stars, and 15K 5 stars. We divide the reviews by stars, and create three datasets: negative (using all of the 1-2 stars), positive (balancing the number of negative reviews using the 5 stars), and neutral (using all of the 3 stars). Table 3 shows our data distribution.

Split	Stars	Num Reviews
POSITIVE	5	7,853
NEUTRAL	3	5,610
NEGATIVE	1-2	7,853

Table 3: Selected Review Data Distribution

3 Learning Patterns for Hyperbole

Our goal is to learn patterns that are highly associated with the extreme positive and negative reviews, and that exemplify strong, expressive language. To automatically learn such patterns, we use the AutoSlog-TS weakly-supervised extraction pattern learner (Riloff, 1996).

AutoSlog-TS uses a set of syntactic templates to learn lexically-grounded patterns. AutoSlog does not require fine-grained labels on training data: all it requires is that the training data be divided into two distinct classes. Here, we run two separate AutoSlog experiments, one in which the classes are POSITIVE compared to NEUTRAL, and the other where the NEGATIVE class is compared to NEUTRAL. We hypothesize that in this way, we can surface the most commonly used patterns from each class that are not necessarily sentiment-related.

AutoSlog applies the Sundance shallow parser (Riloff and Phillips, 2004) to each sentence of each review, finds all possible matches for its syntactic templates, and then instantiates the syntactic templates with the words in the sentence to produce a specific lexico-syntactic expression. Most importantly, it uses the labels associated with the data to compute statistics for how frequently each pattern occurs in each class. Thus, for each pattern p, we learn the P(POSITIVE/NEGATIVE| p), the P(NEUTRAL| p), and the pattern's frequency.

Table 4 shows examples of the patterns we learn and sample instantiations, with their respective frequency (F) and probabilities (P). In the pattern template column of Table 4, PassVP refers to passive voice verb phrases (VPs), ActVP refers to active voice VPs, InfVP refers to infinitive VPs, and AuxVP refers to VPs where the main verb is a form of *to be* or *to have*. Subjects (subj), direct objects (dobj), noun phrases (np), and possessives (genitives) can also be extracted by the patterns. Because we are particularly interested in descriptive patterns, we also use ngram pattern templates, `AdjAdj`, `AdvAdj`, `AdvAdvAdj`, as in related work (Oraby et al., 2015, 2016).

Our goal is to find highly reliable patterns without sacrificing linguistic variation. Current statistical methods for training NLG engines typically eliminate linguistic variability by seeking to learn standard, more generic patterns that occur frequently in the data (Liu et al., 2016; Nayak et al., 2017). Since this phase of our work aims to vastly expand the amount of linguistic variation possible, we select instantiations that have a frequency of at least 3, and a probability of at least 0.75 association with the respective class (Oraby et al., 2015, 2016). We hypothesize that patterns that occur at least 3 times should be fairly reliable, and those that have at least a 75% probability of being associated with the positive or negative class should be distinctive. Using these filters, we learn 8,320 positive adjective patterns, and 7,839 negative adjective patterns.

F	P	Pattern Template	Example Pattern
Positive			
40	1.0	<subj>ActInfVP Dobj	<subj> wait come
19	1.0	ActVP Prep <Np>	tucked in <Np>
54	0.9	AdjAdj	hands down
30	0.9	<subj>ActVP Dobj	<subj> worth wait
20	0.9	NpPrep <Np>	screaming for
10	0.9	<subj> AuxVP Adj	<subj> be scrumptious
416	0.8	AdjNoun	great food
16	0.8	PassVP Prep NP	addicted to
113	0.7	AdvAdj	very fresh
4	0.7	AdjNoun	go-to restaurant
Negative			
17	1.0	<subj> AuxVP Adj	<subj> be impossible
13	1.0	AdjNoun	negative stars
12	1.0	<subj> ActVP Dobj	<subj> got poisoning
23	0.9	AdjNoun	no sense
134	0.8	<subj> AuxVP Adj	<subj> be awful
26	0.8	<subj> AuxVP Adj	<subj> be rubbery
19	0.8	<subj> ActVP	<subj> not waste
107	0.8	AdjNoun	poor service
100	0.8	AdjNoun	no way
201	0.8	<subj> AuxVP Adj	<subj> not be back

Table 4: Examples of Pattern Templates in AutoSlog-TS and Instantiations by Class

We also observe that patterns learned using stricter thresholds (for example, frequency of at least 10 and probability of at least 0.9) also gives us useful patterns, and note that we can use the frequencies and probabilities in our future rank task. For larger coverage, we experiment with our less restrictive thresholds in the current work.

4 Designing Review Templates

To make use of the descriptive adjective patterns we learned, we needed to first identify what entities each of the patterns describes. To do this, we aggregate lexicons for each of five important restaurant entities: *restaurant-type, cuisine, food, service,* and *staff* using Wikipedia[2] and DBpedia[3]. We end up with 14 items for *restaurant-types* (e.g. "cafe"), 45 for *cuisines* (e.g. "Italian"), 4,913 for *foods and ingredients* (e.g. "sushi"), 12 for *staff* (e.g. "waiters"), and 2 for *service* (e.g. "customer service").

4.1 Basic Templates

To construct the most basic set of templates, we use simple relationships between adjectives and the entities they describe to define a set of sentences with entity slots, i.e. *"They had [adj] (entity).", "The (entity) was|is [adj].", "The (entity)*

[2] https://www.wikipedia.org/

[3] http://wiki.dbpedia.org/

looked|tasted [adj].” We use basic lists of adjectives commonly found in reviews for these baseline templates. To vary the templates, we alternate between using only simple sentences, and sometimes combine related entities into more complex sentences (e.g. *service* and *staff*, or *restaurant-type* and *cuisine*).

4.2 Hyperbolic Templates

For our hyperbolic templates, we replace the standard adjectives in the basic templates with adjective patterns learned from the restaurant reviews. To select appropriate adjectives patterns for replacement in each basic template, we first delexicalize the sentences that instantiate our learned adjective patterns for each class, and create sets of (entity, adjective pattern) pairs based on the relationship between the adjective and the entity (“is”, “was”, “tasted”, etc.), as above. Using this method, we collect 37 restaurant, 30 cuisine, 247 food, 45 service, and 56 staff patterns for positive and 18 restaurant, 9 cuisine, 221 food, 75 service, and 61 staff patterns for negative. Table 5 shows example patterns in each class for the food and staff entity types.

Positive	**Negative**
INSANELY GOOD	ALMOST RAW
SIMPLY PERFECT	VERY FATTY
RIDICULOUSLY GOOD	PREVIOUSLY FROZEN
ALSO INCREDIBLE	COMICALLY BAD
MY FAV	ABSOLUTELY AWFUL
PERFECTLY CRISP	NOT PALATABLE
DEFINITELY UNIQUE	FAIRLY TASTELESS
ALWAYS SO FRESH	PRETTY GENERIC
JUST PHENOMENAL	SO MEDIOCRE
SO DECADENT	SO BLAND
HIGHLY ADDICTIVE	STILL RAW
CONSISTENTLY GREAT	BARELY WARM
WOW AMAZING	PREPACKAGED FROZEN
PERFECT LITTLE	MOST PATHETIC
EXPERTLY PREPARED	SICKLY SWEET
FRESHLY BAKED	LUKE WARM

(a) Sample Learned Adjective Patterns for Foods

Positive	**Negative**
SUPER HELPFUL	NOT APOLOGETIC
INCREDIBLY FRIENDLY	NOT KNOWLEDGEABLE
SUPER NICE	VERY RUDE
VERY PERSONABLE	TOO BUSY
SO GOOD	FRIENDLY ENOUGH
SO GRACIOUS	JUST HORRIBLE
VERY KNOWLEDGEABLE	NOT ATTENTIVE
SO KIND	VERY PUSH
EXTREMELY PROFESSIONAL	MORE INTERESTED
ALSO FABULOUS	TOO LAZY
EVEN BETTER	EVEN WORSE
STILL AWESOME	EVERY SINGLE
ALWAYS WARM	VERY POOR
ALWAYS ATTENTIVE	SO FEW
ABSOLUTELY BEST	STILL NO
OUR SWEET	VERY UNHAPPY

(b) Sample Learned Adjective Patterns for Staff

Table 5: Sample Learned Adjective Patterns

4.3 Creative Templates

Finally, for our creative templates, we sample from our set of delexicalized sentences for each entity type, as long as they:

- contain a single AutoSlog adjective pattern
- contain a single identifiable entity type
- are between 5-15 words long

We enforce these limitations to gather simple sentences that are short enough to templatize. Thus, we end up with sentence templates for each entity type for both the positive and negative classes, collecting 146 restaurant, 61 cuisine, 743 food, 90 service, and 144 staff patterns for positive and 45 restaurant, 12 cuisine, 480 food, 126 service, and 89 staff patterns for negative. Table 6 shows examples of our templatized sentences for the positive and negative classes, with their AutoSlog-TS adjective patterns between brackets, and capitalized subject extractions when applicable. To construct a full review of a certain polarity, we randomly select a sentence from the sets for each entity type.

We hypothesized that the creative templates would optimize stylistic variation and hence interestingness, but that they would also include cases that would require further refinement, or perhaps elimination by a subsequent ranking phase. Since our focus here is on overgeneration, we include these and evaluate their quality. Table 7 shows examples of each template type we create.

5 Evaluating Template Styles

In order to evaluate our template variations, we choose to focus on three particular criteria: *convincingness, interestingness,* and *naturalness*. We evaluate *convincingness* because creative language such as hyperbole is often used in persuasive language, along with other figurative forms (Kreuz and Roberts, 1995). *Naturalness* is an important concern in generation, so we are also interested in the comparison between the perceived naturalness of each variation style, and we hypothesize that *interestingness* would increase as we used

Entity	Template
Positive	
RESTAURANT	By [FAR MY] favorite <RESTAURANT_ENTITY> I HAVE EVER been to in my life .
CUISINE	Wow what a great [LITTLE <CUISINE_ENTITY>] joint !
FOOD	The <FOOD_ENTITY> is not cheap , but [WELL WORTH] it.
SERVICE	The <SERVICE_ENTITY> is [ALWAYS FRIENDLY] and fast .
STAFF	<STAFF_ENTITY> was [EXTREMELY HELPFUL] and knowledgeable and was on top of everything.
Negative	
RESTAURANT	I was appalled by the experience and will [NOT FREQUENT] this <RESTAURANT_ENTITY> ever again.
CUISINE	[ITS YOUR] typical <CUISINE_ENTITY> buffet , nothing to rave about .
FOOD	<FOOD_ENTITY> smelled [VERY BAD] and tasted worse .
SERVICE	We waited another 5 minutes , [STILL NO] <SERVICE_ENTITY> .
STAFF	I went with 5 friends and our <STAFF_ENTITY> was [REALLY RUDE] .

Table 6: Examples of Learned Creative Sentence Templates by Entity and Polarity

BASIC	The bar is beautiful. They had authentic japanese cuisine. The udon looked excellent. The hosts is dedicated. They had reliable customer service.
HYPERBOLIC	The bar is also very fresh. They had delicious authentic japanese cuisine. The udon looked so delicious. The hosts is also very friendly. They had such amazing customer service.
CREATIVE	This is by far my favorite bar in town. plus there is a great japanese cuisine grocery store that has tons of stuff. The udon is always fresh, delicious and made to order. Hosts was super friendly, looking forward to coming back and trying more items. The customer service is great and the employees are always super nice!

Table 7: Examples of Instantiated Positive Review Variations

more content from organic reviews in our HYPERBOLIC and CREATIVE templates.

To create an evaluation dataset, we instantiate each template type with entities from a hypothetical MR in one of seven popular cuisine types to standardize the content, as illustrated in Table 7. For example, sample slot values could be: {RESTAURANT[BAR], CUISINE[JAPANESE], FOOD[UDON], SERVICE[CUSTOMER SERVICE], STAFF[HOSTS]}.

Our objective is to evaluate whether we can improve upon vanilla-style hand-crafted templates for restaurant reviews by utilizing in hyperbolic and creative elements of organic reviews that we harvest. We set up an annotation experiment on Amazon Mechanical Turk[4], where each Human Intelligence Task (HIT) presents Turkers with a sample of our three review variations, all of the same polarity and instantiated with the same entities. Turkers are asked to judge the reviews based on three criteria: *convincingness* (Do you believe the opinion given?), *interestingness* (Is the review engaging?), and *naturalness* (Is the review coherent?). Turkers are asked to rate each review on a three point scale (*high, medium* and *low*) for each criteria. We release 200 variation triples (100 per polarity class) and ask for five judgements per HIT, tagging a review with a quality if the majority of annotators agree on it (i.e. 3 or more Turkers). Average agreement for individual Turkers with the majority is above 73%.

Figure 1 shows the distribution of *high, medium,* and *low* scores for each of the variation types for each criterion. From the results, we observe that for all criteria, the CREATIVE class has the highest distribution of *high* majority votes. Interestingly, although we hypothesized that the HYPERBOLIC reviews would be better received than the BASIC reviews, we observe that in fact the BASIC reviews receive more *high* votes on *convincingness*. We note that for the future ranking, more context information is necessary when selecting appropriate hyperbolic patterns with which to modify the BASIC reviews. For example, if a learned pattern is OTHER AMAZING, the pattern should only be used when a set of items are being described, and not stand-alone. Similarly, the BASIC reviews are also more *natural* than the HYPERBOLIC ones, although both variation types score very similar percentages for *medium* scores.

For the creative reviews, a crucial next step for ranking is to consider context and develop heuristics for finding the most appropriate entities for lexicalization. For example, for very

[4] https://www.mturk.com/

specific creative templates such as: *"I also got one that HAD NOT been separated , so it was [JUST HALF] of a* <FOOD_ENTITY> *."*, or *"The* <FOOD_ENTITY> *were similarly a mix of nearly raw to overly crisp."*, it is necessary to select food items similar to the original instantiations, or to characterize and classify entities based on specific properties.

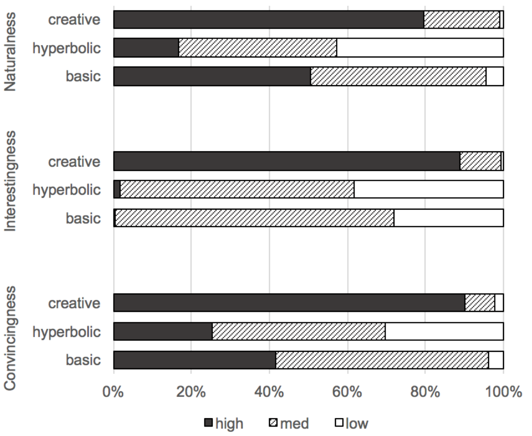

Figure 1: Distribution of Template Variations by Evaluation Criteria

Given the high appeal of the CREATIVE reviews on all counts, we are interested in more closely exploring examples in the data. Table 8 shows two examples of CREATIVE reviews: one that received *high* scores on all criteria, and one that received majority (no creative review received all *lows*). It is clear that the biggest disconnect in the low-scoring *creative* review is the coherence between sentences, which as an important next step to consider as future work given the proof-of-concept presented here. We also note that we can also improve the fully high-scoring review by fixing grammatical errors and applying more informed content selection.

To get a better sense of how grammatically correct the review template variations are, we conduct another evaluation study where we present Turkers with the same set of reviews, and ask them to rate each review based on the content (checking subject-verb agreement, plurality, tense, etc.). Similar to the previous study, we gather 5 judgements for each set of three variations, and aggregate results using majority vote. Average agreement for individual Turkers with the majority in this task is above 80%, higher than the more subjective study on convincingness, interestingness, and naturalness.

ALL HIGH	It is one of my favorite cafe in las vegas. Thank you irma for your amazing mediterranean cuisine cooking! I am always amazed at how fast my falafel arrives. Victor the owner was super nice and cordial our hosts norma was also. Always a great place to go and service is always amazing!
MOSTLY LOW	It's just too bad that the bar itself is not better. Very bad american cuisine.... Guess what came on top of my hotdog? I took my family there for father's day and the hosts was so rude. 555 pm still no customer service.

Table 8: Example of *High* and *Low* Rated Creative Reviews

Figure 2 shows the results of the study. We find that for all three variations, the *med* class receives the majority of the votes, but that the BASIC reviews are the most grammatically correct (since the templates are designed, not harvested). Similarly, the HYPERBOLIC reviews have the largest percentage of *low* scores, since their creation involves modifying templates with learned adjectives. Ranking the best patterns/sentences to use will allow us to improve the grammatical coherence of the templatized utterances for the HYPERBOLIC and CREATIVE classes.

To better understand the linguistic characteristics of the creative reviews by class, we run the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et al., 2015) on the full set of 100 POSITIVE and 100 NEGATIVE *creative* reviews. When comparing the linguistic categories for each class, we find that the difference between the POSITIVE and NEGATIVE reviews are significant ($p < 0.05$, t-test) for many of the categories. Table 9 shows some of the most interesting categories[5].

On average, the POSITIVE templates are char-

[5] All of the categories are statistically significant, and are shown in order of most to least significant.

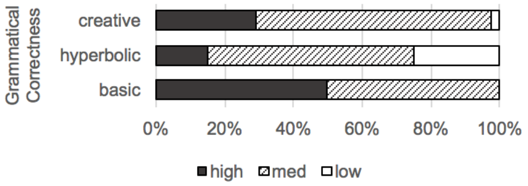

Figure 2: Distribution of Votes on Template Variation Grammar

Positive	Negative
AFFECTIVE PROC.	DIFFERENTIATION
EXCLAMATIONS	RISK
FRIENDS	1^{st} PERSON PLURAL
1^{st} PERSON SINGULAR	ANXIETY
ACHIEVE	ADVERBS
CERTAINTY	ANGER
BIOLOGICAL PROC.	SOCIAL PROC.
INGESTIONS	2^{nd} PERSON
INSIGHT	MOTION
REWARD	COGNITIVE PROC.

Table 9: Statistically Significant LIWC Categories by Polarity

acterized by word classes that exemplify achievement (e.g. *"even better", "champion"*) and certainty (e.g. *"always excellent", "absolutely amazing"*, and *"definitely my go-to place"*). As well as 1^{st} person statements relating to use of the senses (affective processes like *"my favorite place to get rice in Las Vegas!"*, biological processes (*"I just had the most amazingly delicious and freshly prepared couscous!"*), and ingestion (*"good, tasty comfort pizza"*).

The negative contains more oppositional language directed at the second person, often as advice (*"you can get a much better pizza elsewhere at far less cost."*), with categories like differentiation (*"but it's not great"*), and strong emotion indicators like anxiety (*"horrible service, finally just left"*) and anger (*"I was so angry that I contacted the restaurant manager"*).

6 Conclusions

In this paper, we show that we can construct convincing, interesting, and natural restaurant review templates by using a data-driven method to harvest highly descriptive sentences from hyperbolic restaurant reviews. We generate three variations of review templates, ranging from very basic, to hyperbolic, to very creative, and show that the creative ones are more appealing to readers than the others. Future work will focus on ranking the candidate sentence templates we harvest to improve review coherence. As we develop better templates, we will evaluate them against baselines from existing NLG systems to guide our generation of more exciting and expressive stylistically varied reviews.

References

Srinivas Bangalore and Owen Rambow. 2000. Exploiting a probabilistic hierarchical model for generation. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pages 42–48.

Laura Cano Mora. 2009. All or nothing: a semantic analysis of hyperbole. *Revista de Lingüística y Lenguas Aplicadas* pages 25–35.

Herbert L. Colston and Shauna B. Keller. 1998. You'll never believe this: Irony and hyperbole in expressing surprise. *Journal of Psycholinguistic Research* 27(4):499–513.

Herbert L. Colston and Jennifer O'Brien. 2000. Contrast and pragmatics in figurative language: Anything understatement can do, irony can do better. *Journal of Pragmatics* 32(11):1557 – 1583.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen F. Hastie, Verena Rieser, and Oliver Lemon. 2014. Cluster-based prediction of user ratings for stylistic surface realisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*. pages 702–711.

Laurence Devillers, Hélène Maynard, Sophie Rosset, Patrick Paroubek, Kevin McTait, Djamel Mostefa, Khalid Choukri, Laurent Charnay, Caroline Bousquet, Nadine Vigouroux, Frédéric Béchet, Laurent Romary, Jean-Yves Antoine, Jeanne Villaneau, Myriam Vergnes, and Jérôme Goulian. 2004. The french MEDIA/EVALDA project: the evaluation of the understanding capability of spoken language dialogue systems. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*.

Ondrej Dusek and Filip Jurcícek. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. *CoRR* abs/1606.05491.

Milica Gasic, Simon Keizer, Francois Mairesse, Jost Schatzmann, Blaise Thomson, Kai Yu, and Steve Young. 2008. Training and evaluation of the his pomdp dialogue system in noise. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, Association for Computational Linguistics, Columbus, Ohio, page 112–119.

Ryuichiro Higashinaka, Marilyn Walker, and Rashmi Prasad. 2007a. Learning to generate naturalistic utterances using reviews in spoken dialogue systems. *ACM Transactions on Speech and Language Processing (TSLP)* .

Ryuichiro Higashinaka, Marilyn A. Walker, and Rashmi Prasad. 2007b. An unsupervised method for learning generation dictionaries for spoken dialogue systems by mining user reviews. *ACM Transactions on Speech and Language Processing* 4(4).

David Howcroft, Crystal Nakatsu, and Michael White. 2013. Enhancing the expression of contrast in the sparky restaurant corpus. In *Proceedings of the 14th European Workshop on Natural Language Generation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 30–39.

Roger J. Kreuz and Richard M. Roberts. 1995. Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and Symbolic Activity* 10(1):21–31.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*. ACL '98, pages 704–710.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. pages 2122–2132.

Francois Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue. In *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL*. pages 496–503.

Francois Mairesse and Marilyn Walker. 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction* pages 1–52.

Michael McCarthy and Ronald Carter. 2004. There's millions of them: hyperbole in everyday conversation. *Journal of Pragmatics* 36(2):149–184.

Neha Nayak, Dilek Hakkani-Tur, Marilyn Walker, and Larry Heck. 2017. To plan or not to plan? discourse planning in slot-value informed sequence to sequence models for language generation. In *Proc. of Interspeech 2017*.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing NLG data: Pictures elicit better data. In *INLG 2016 - Proceedings of the Ninth International Natural Language Generation Conference, September 5-8, 2016, Edinburgh, UK*. pages 265–273.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn A. Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*. pages 31–41.

Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. 2015. And that's a fact: Distinguishing factual and emotional argumentation in online dialogue. In *Proceedings of the 2nd Workshop on Argumentation Mining at NAACL 2015*. pages 116–126.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Joseph Polifroni, Lynette Hirschman, Stephanie Seneff, and Victor Zue. 1992. Experiments in evaluating interactive spoken language systems. In *Proceedings of the Workshop on Speech and Natural Language*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '91, pages 28–33.

Owen Rambow, Monica Rogati, and Marilyn Walker. 2001. Evaluating a trainable sentence planner for a spoken dialogue travel system. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*. pages 1044–1049.

Ellen Riloff and William Phillips. 2004. An introduction to the sundance and autoslog systems. Technical report, Technical Report UUCS-04-015, School of Computing, University of Utah.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialogue systems. In *Meeting of the Association for Computational Linguistics*.

Marilyn Walker and Owen Rambow. 2002. Spoken language generation. *Computer Speech & Language* 16(3):273 – 281. Spoken Language Generation.

Marilyn Walker, Amanda Stent, Francois Mairesse, and Rashmi Prasad. 2007. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research (JAIR)* 30:413–456.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *CoRR* abs/1508.01745.

Steve Whittaker, Marilyn Walker, and Johanna Moore. 2002. Fish or fowl: A wizard of oz evaluation of dialogue strategies in the restaurant domain. In *Language Resources and Evaluation Conference*.

Is writing style predictive of scientific fraud?

Chloé Braud and **Anders Søgaard**
CoAStaL DIKU
University of Copenhagen
University Park 5, 2100 Copenhagen
chloe.braud@gmail.com soegaard@di.ku.dk

Abstract

The problem of detecting scientific fraud using machine learning was recently introduced, with initial, positive results from a model taking into account various general indicators. The results seem to suggest that writing style is predictive of scientific fraud. We revisit these initial experiments, and show that the leave-one-out testing procedure they used likely leads to a slight over-estimate of the predictability, but also that simple models can outperform their proposed model by some margin. We go on to explore more abstract linguistic features, such as linguistic complexity and discourse structure, only to obtain negative results. Upon analyzing our models, we do see some interesting patterns, though: Scientific fraud, for examples, contains less comparison, as well as different types of hedging and ways of presenting logical reasoning.

1 Introduction

Cases of scientific misconduct are identified every year. Scientific papers are retracted because of errors, or for suspected fraud, ranging from plagiarism and minor manipulations to faking the data and disguising the results. It has been shown that, however, among the retracted articles indexed in PubMed, only 21.3% are retracted due to error, while 67.4% were removed due to misconduct, among which suspected fraud amounts to 43.4%, the others being due to duplicate publications or plagiarism (Fang et al., 2012).

In a recent paper, Markowitz and Hancock (2015) proposed the first analysis of writing style in fraudulent papers across authors and disciplines. They approached the question of whether these authors have a specific writing style, from a psychological perspective. They found that these papers exhibit a higher rate of jargon, make a higher use of references, and have a lower readability rate, suggesting that the authors try to obfuscate their writing, making them harder to read and analyze. They report classification results using a leave-one-out strategy over the dataset, with a classification accuracy of 57.2%. As suggested in the paper, we propose to improve this performance by evaluating different classification models.

In this paper, we first show that much better results can be obtained using a simple bag-of-words representation and Logistic Regression. Our best model is a syntax-enhanced trigram-model. We also show that the leave-one-out strategy used by the authors leads to an over-estimation of model precision, and we report new results based on a more robust strategy, taking into account the low number of instance available; namely a *nested* cross-validation (Varma and Simon, 2006; Scheffer, 1999). We also considered semantic and discourse features, but we did not observe improvements with such features.

Of course, that a bag-of-words model outperforms a model based on psychologically motivated features, may simply be the result of overfitting. We present an extensive feature analysis to validate our models, as well as to test psychologically motivated hypotheses from the literature.

Contributions (i) We present a simple model with high accuracy, and show that it implicitly captures the previously-proposed psychologically-motivated features. (ii) We show that adding semantics and discourse features does not lead to improvements. (iii) On the other hand, our feature analysis suggests that the models *do* learn to focus on concepts that are intuitively related to scientific

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 37–42
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics