Linguistically-Informed Specificity and Semantic Plausibility for Dialogue Generation

Wei-Jen Ko[1] **Greg Durrett**[1] **Junyi Jessy Li**[2]
[1] Department of Computer Science
[2] Department of Linguistics
The University of Texas at Austin
wjko@utexas.edu, gdurrett@cs.utexas.edu, jessy@austin.utexas.edu

Abstract

Sequence-to-sequence models for open-domain dialogue generation tend to favor generic, uninformative responses. Past work has focused on word frequency-based approaches to improving specificity, such as penalizing responses with only common words. In this work, we examine whether specificity is solely a frequency-related notion and find that more linguistically-driven specificity measures are better suited to improving response informativeness. However, we find that forcing a sequence-to-sequence model to be more specific can expose a host of other problems in the responses, including flawed discourse and implausible semantics. We rerank our model's outputs using externally-trained classifiers targeting each of these identified factors. Experiments show that our final model using linguistically motivated specificity and plausibility reranking improves the informativeness, reasonableness, and grammatically of responses.

1 Introduction

Since the pioneering work in machine translation (Sutskever et al., 2014), sequence-to-sequence (SEQ2SEQ) models have led much recent progress in open-domain dialogue generation, especially single-turn generation where the input is a prompt and the output is a response. However, SEQ2SEQ methods are known to favor universal responses, e.g., "I don't know what you are talking about" (Sordoni et al., 2015; Serban et al., 2016; Li et al., 2016a). These responses tend to be "safe" responses to many input queries, yet they usually fail to provide useful information.

One promising line of research tackling this issue is to improve the specificity of responses, building on the intuition that generic responses frequently appear in the training data or consist of frequent words (Yao et al., 2016; Zhang et al., 2018b; Liu et al., 2018). However, past work in sentence specificity—the "quality of belonging or relating uniquely to a particular subject"[1]—has shown that word frequency is only one aspect of specificity, and that specificity involves a wide range of phenomena including word usage, sentence structure (Louis and Nenkova, 2011; Li and Nenkova, 2015; Lugini and Litman, 2017) and discourse context (Dixon, 1987; Lassonde and O'Brien, 2009). Frequency-based specificity also does not exactly capture "the amount of information" as an information-theoretic concept. Hence, in dialogue generation, we can potentially make progress by incorporating more linguistically driven measures of specificity, as opposed to relying solely on frequency.

We present a sequence-to-sequence dialogue model that factors out specificity and explicitly conditions on it when generating a response. The decoder takes as input categorized values of several specificity metrics, embeds them, and uses them at each stage of decoding. During training, the model can learn to associate different specificity levels with different types of responses. At test time, we set the specificity level to its maximum value to force specific responses, which we found to be most beneficial. We integrate linguistic (Ko et al., 2019), information-theoretic, and frequency-based specificity metrics to better understand their roles in guiding response generation.

The second component of our model is designed to make the more specific responses more *semantically plausible*. In particular, we found that forcing a SEQ2SEQ model to be more specific exposes problems with plausibility as illustrated in Table 1. As sentences become more specific and contain more information, intra-response consistency

[1]Definition from the Oxford Dictionary

Proceedings of NAACL-HLT 2019, pages 3456–3466
Minneapolis, Minnesota, June 2 - June 7, 2019.

Conflicting	i understand. i am not sure if i can afford a babysitter, i am a millionaire
Wrong connective	i am an animal phobic, but i do not like animals
Wrong pronoun	my mom was a social worker, he was an osteopath.
Wrong noun	cool. i work at a non profit organization that sells the holocaust.
Repeating	my favorite food is italian, but i also love italian food, especially italian food.

Table 1: Examples of different types of implausible responses on the PersonaChat dataset generated from our system that maximizes specificity only.

problems become evident, making the overall response implausible or unreasonable in real life. Our inspection discovered that ∼30% of specific responses suffer from a range of problems from semantic incompatibility to flawed discourse. To improve the plausibility of responses, we propose a reranking method based on four external classifiers, each targeting a separate aspect of linguistic plausibility. These classifiers are learned on synthetically generated examples, and at test time their responses are used to rerank proposed responses and mitigate the targeted issues.

Using both automatic and human evaluation, we find that linguistic-based specificity is more suitable than frequency-based specificity for generating informative and topically relevant responses, and learning from different types of specificity metrics leads to further improvement. Our plausibility reranking method not only successfully improved the semantic plausibility of responses, but also improved their informativeness, relevance, and grammaticality.

Our system is available at `https://git.io/fjkDd`.

2 Related work

Generic responses is a recognized problem in dialogue generation. Li et al. (2016a) maximized mutual information in decoding or reranking, which practically looks like penalizing responses that are common under a language model. Zhou et al. (2017) promoted diversity by training latent embeddings to represent different response mechanisms. Shao et al. (2017) trained and reranked responses segment by segment with a glimpse model to inject diversity. Another angle is to promote prompt-response coherence using techniques such as LDA (Baheti et al., 2018; Xing et al., 2017). Cosine similarity between prompt and response has also been used for coherence (Xu et al., 2018b; Baheti et al., 2018). Wu et al. (2018) learn a small vocabulary of words that may be relevant during decoding and generates responses with this vocabulary.

Several works tackle the problem by directly controlling response specificity in terms of word and response frequency. IDF and response frequency have been used as rewards in reinforcement learning (Yao et al., 2016; Li et al., 2016d). Some methods adjusted sample weights in the training data, using a dual encoding model (Lison and Bibauw, 2017) or sentence length and frequency in the corpus (Liu et al., 2018). Zhang et al. (2018b) proposed a Gaussian mixture model using frequency-based specificity values. Their approach involves ensembling the context probability and a specificity probability, whereas our approach conditions on both in a single model.

Prediction of **sentence specificity** following the dictionary definition and pragmatically cast as "level of detail" was first proposed by Louis and Nenkova (2011), who related specificity to discourse relations. Sentence specificity predictors have since been developed (Louis and Nenkova, 2011; Li and Nenkova, 2015; Lugini and Litman, 2017; Ko et al., 2019). Insights from these feature-rich systems and hand-code analysis (Li et al., 2016e) showed that sentence specificity encompasses multiple phenomena, including referring expressions, concreteness of concepts, gradable adjectives, subjectivity and syntactic structure.

Researchers have noticed that distributional semantics largely fail to capture **semantic plausibility**, especially in terms of discrete properties (e.g., negation) (Kruszewski et al., 2016) and physical properties (Wang et al., 2018). Kruszewski et al. (2016) created a dataset building on synthetically generated sentences for negation plausibility.

Methodology-wise, Li et al. (2016b) trained embeddings for different speakers jointly with the dialogue context. Huang et al. (2018) learned embeddings of emotions; we learn embeddings of specificity metrics. Targeting multiple factors this way is broadly similar to the approach of Holtzman et al. (2018), who used multiple cooperative discriminators to model repetition, entailment, rel-

evance, and lexical style in generation. Our approach additionally leverages synthetic synthetic sentences targeting a range of plausibility issues and trains discriminators for reranking.

3 Generating specific responses

Our main framework (Figure 1) is an attention-based SEQ2SEQ model (Section 3.1) augmented with the ability to jointly learn embeddings from a target metric (e.g., specificity) with the response (Section 3.2). We then integrate frequency-based, information-theoretic and linguistic notions of specificity (Section 3.3) as well as coherence (Section 3.4).

3.1 Base framework

Our model is based on a SEQ2SEQ model (Sutskever et al., 2014) consisting of an encoder and decoder, both of which are LSTMs (Hochreiter and Schmidhuber, 1997). We apply attention (Bahdanau et al., 2015) on the decoder. The encoder LSTM takes word embeddings x_i in the prompt sentence as input. The hidden layer and cell state of the decoder are initialized with the final encoder states. During training, the decoder takes the embedding of the previous word in the gold response as input; during testing, it uses the previous generated word. We denote both as y_{i-1}:

$$h_i^d, c_i^d = LSTM(y_{i-1}, [\hat{h}_i^d; c_{i-1}^d]) \quad (1)$$

where $\hat{h}_i^d$ is the output of the attention mechanism, given the decoder hidden state.

During training, we minimize the negative log likelihood of responses Y given prompts X.

3.2 Conditioning on specificity

In the base model, uninformative responses are preferred partially because these are common in the training data. We want to be able to fit the training data while at the same time recognizing that we do not want to generate such responses at test time. Our approach, shown in Figure 1, involves conditioning on an explicit specificity level during both training and test time. This explicit conditioning allows us to model specificity orthogonally to response content, so we can control it at test time. We represent specificity as a collection of real valued metrics that can be estimated for each sentence independently of the dialogue system. To direct the model to generate more specific responses from multiple specificity metrics,

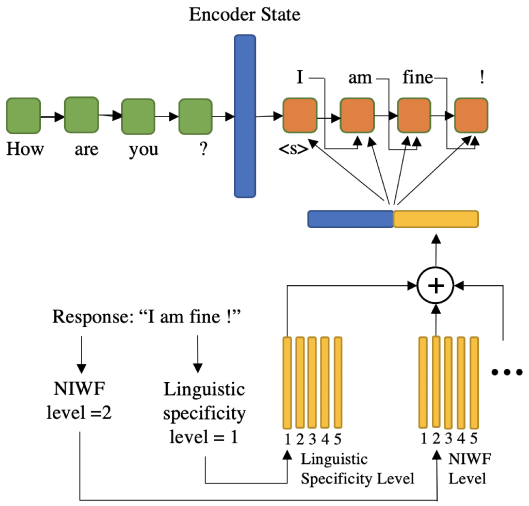

Figure 1: Structure of our model. The decoder explicitly conditions on embeddings of various specificity measures. At train time, these factor out specificity from generation; at test time, maxing these out encourages the model to generate specific responses. NIWF is defined in Section 3.3.

we learn embeddings of various specificity levels for each metric jointly with the model.

In particular, for each metric m, we rank the responses in the training data according to that metric and divide it into $K = 5$ levels of equal size. For each level, we learn an embedding e_k^m, $k \in \{1, 2, ...K\}$. During training, for each sentence pair in the training set, the response is classified to level l_m for metric m. We take the sum of embeddings across all metrics $e = \sum_{m=1}^{N} e_{l_m}^m$ and feed it into the decoder at every time step, where N is the number of metrics. The decoder becomes

$$h_i^d, c_i^d = LSTM(y_{i-1}, [\hat{h}_i^d; c_{i-1}^d; e]) \quad (2)$$

During testing, we specify a level for each metric and calculate e based on those levels. In practice, the level of specificity varies with the larger context of dialogue discourse, however for the purpose of avoiding generic responses and improving specificity in single-turn dialogue generation, and examining various metrics of specificity, we use the level that maximizes specificity at test time (which we show in Section 5.3 is better the uninformative "median" level).[2]

[2]For the purposes of this work, we want an agent that is highly specific and keeps the conversation going. Learning the ideal specificity for a given response is something we leave for future work.

3.3 Specificity metrics

Normalized inverse word frequency (NIWF) Used in Zhang et al. (2018b), NIWF is the maximum of the Inverse Word Frequency (IWF) of all the words in a response, normalized to 0-1:

$$\max(IWF) = \max\left(\frac{\log(1+|Y|)}{f_w}\right) \quad (3)$$

where f_w denotes the number of responses in the corpus that contain the word w, and $|Y|$ is the number of responses in the corpus. Taking a maximum reflects the assumption that a response is specific as long as it has at least some infrequent word.

Perplexity per word (PPW) Perplexity is the exponentiation of the entropy, which estimates the expected number of bits required to encode the sentence (Brown et al., 1992; Goodman, 2001). Thus perplexity is a direct measure of the amount of information in the sentence in information theory; it has also been used as a measure of linguistic complexity (Gorin et al., 2000). To compute perplexity, we train a neural language model (Mikolov et al., 2011) on all gold responses and calculate cross-entropy of each sentence. To represent the amount of information per-token and to prevent the model to simply generate long sentences, we normalize perplexity by sentence length.

Linguistically-informed specificity We use the system developed by Ko et al. (2019), which estimates specificity as a real value. This system adopts a pragmatic notion of specificity—level of details in text—that is originally derived using sentence pairs connected via the INSTANTIATION discourse relation (Louis and Nenkova, 2011). With this relation, one sentence explains in further detail of the content in the other; the explanatory sentence is shown to demonstrate properties of specificity towards particular concepts, entities and objects, while the other sentence is much more general (Li and Nenkova, 2016). We use this particular system since other specificity predictors are trained on news with binary specificity labels (Li and Nenkova, 2015). Ko et al. (2019) is an unsupervised domain adaptation system that predicts continuous specificity values, and was evaluated to be close to human judgments across several domains. We retrain their system using the gold responses in our data as unlabeled sentences in the unsupervised domain adaptation component.

3.4 Coherence

Prior work has shown that the universal response problem can be mitigated by improving the coherence between prompt and response (Zhang et al., 2018a; Xu et al., 2018b; Baheti et al., 2018). We introduce two methods to improve coherence upon the base model, and analyze specificity on top.

For better interactions between decoder embeddings and the prompt, we feed the final encoder state into every time step of the decoder, instead of only the first token. Thus the decoder becomes

$$h_i^d, c_i^d = LSTM(y_{i-1}, [\hat{h_i^d}; c_{i-1}^d; e; h_f]). \quad (4)$$

Furthermore, Zhang et al. (2018a) showed that responses ranked higher by humans are more similar to the prompt sentence vector. Thus we compute the cosine similarity between input and response representations. This is computed by the weighted average of all word embeddings in the sentence, where the weight of each word is its inverse document frequency. Our model additionally conditions on an embedding of this measure so that coherence is factored out in our model as well as specificity. During testing, we condition on the highest level of our similarity metric in order to generate maximally coherent responses (Xu et al., 2018b).

4 Semantic plausibility

While injecting specificity encourages the model to generate more specific responses, we discovered that it exposes a series of issues that together, severely impact the semantic plausibility of generated responses. This is the case even when responses are considered independently without the prompt context. To have a better understanding of the problem, we first present manual analysis on generated responses with improved specificity. We then present a reranking method to improve the semantic plausibility of responses.

4.1 Data analysis

We manually inspected 200 responses generated from our full model on the PersonaChat dataset (Zhang et al., 2018c). We evaluated the responses independent of the input prompt and found that ∼33% of the sentences are semantically implausible; some of them shown in Table 1.

We found three major types of errors. The most common type is a wrong word that is not compatible with the context, making the phrase unreasonable (*cool . i work at a non profit organization that*

sells the ***holocaust***), meaningless (*i like to* ***dance*** *battles*), or unnatural (*yeah , but i am more of a* ***game*** *worm . i am a pro ball player)*. These make up about 45% of the implausible cases.

About 30% of the problematic sentences contain incompatible phrases. Different phrases in the response are contradictory (*i understand. i am not sure if i can afford a babysitter, i am a millionaire*) or repetitive (*my favorite food is italian, but i also love italian food, especially italian food.*).

The third problem (∼15%) is that phrases are connected by a wrong discourse connective (*i am an animal phobic,* ***but*** *i do not like animals*). This and the previous problem reveal that even when the model generates sensible phrases, proper discourse relations between them are not captured.

Other notable errors include cohesion, such as wrong determiners or pronouns (*my mom was a social worker,* ***he*** *was an osteopath.*) and inappropriate prepositional phrases (*hello , i am winding down* ***to*** *the morning .*)

This semantic implausibility may come from two sources. First, since specific responses tend to be longer, it is easier to have internal consistency issues where parts of the sentence are incompatible with each other. Second, regardless of the specificity metric, word frequency in specific responses tend to be lower than that in generic responses. Learning meaningful representations for infrequent words is a known challenge (Gong et al., 2018) hence low-quality representations may increase the probability of the sentence being implausible.

4.2 Reranking

To mitigate semantic plausibility issues, we propose a reranking method so that more plausible sentences are ranked higher among the candidates. We use classifiers targeting various types of errors using synthetically generated data. Specifically, we train four classifiers that distinguish true response sentences from the dataset and negative sentences we create that reflect a specific type of semantic implausibility:

Phrase compatibility: We split all the training data into phrases by splitting sentences on punctuation or discourse connectives. To create a negative sentence given a gold response, we pick a random phrase in the true response and replace it with a random phrase in another random true response.

Content word plausibility: We replace a randomly selected content word (noun, verb, adjective, adverb) in the gold response with another random word with the same part-of-speech in the training set.

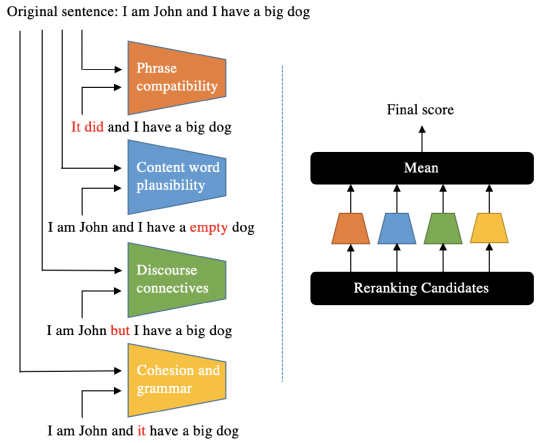

Figure 2: Reranking models to encourage plausibility. Four types of errors are synthetically applied to the data and classifiers are trained to differentiate each transformed sentence from the original. The mean score under these classifiers is then used as a feature to rerank system outputs.

Discourse connectives: We replace a discourse connective in the gold response (if one exists) with a random connective.

Cohesion and grammar: We replace a randomly selected function word in the gold response with another random function word of the same part-of-speech. For pronouns and determiners, these negative sentences would likely be incohesive; with other word categories such as prepositions, this will target grammatically.

One word or phrase is replaced in each synthetic sentence. We train one classifier $\theta_j, j \in \{1,2,3,4\}$ for each of the categories above.[3] The classifiers take word embeddings as input and predict if the response is real or generated. Each classifier consists of a bi-directional LSTM with a projection layer and max pooling (Conneau et al., 2017), followed by 3 fully connected layers. The posterior probabilities of these classifiers reflect how confident the classifiers are that the sentence is synthetic and prone to be implausible, hence we prefer sentences with lower posterior probabilities. During reranking, we feed each candidate sentence c into the classifiers and aggregate the posterior probabilities from these classifiers by taking the

[3]We compare with using one classifier lumping all negative sentences in the experiments.

mean $\frac{1}{4}\sum_{k=1}^{4} P(synthetic|c, \theta_k)$.

At test time, to encourage diversity, we repeat inference multiple times to generate different candidate sentences, and each time dropout is applied to different nodes in the network. Compared with diverse decoding (Li et al., 2016c), we observed during development that sentences generated by different dropouts tend to have diverse semantics (hence more likely to have different plausibility levels). On the contrary, sentences from diversity decoding often have similar structure and phrases across candidates. We also experimented with reinforcement learning, using policy gradient with the reranking scores as reward. However, during development, we observed that this method produced shorter, less informative sentences compared to reranking.

5 Experiments

5.1 Evaluation metrics

Automatic evaluation of dialogue generation systems is a known challenge. Prior work has shown that commonly used metrics for overall quality in other generation tasks such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and perplexity have poor correlations with human judgment (Liu et al., 2016; Tao et al., 2018)[4] or are model-dependent (Liu et al., 2016). Therefore, we adopt several metrics that evaluate multiple aspects of responses, and also conduct human evaluation for each result we present.

We use the following automatic evaluation metrics: (1) **distinct-1** and **distinct-2** (Li et al., 2016a), which evaluates response *diversity*. They respectively calculate the number of distinct unigrams and bigrams, divided by the total number of words in all responses; (2) linguistically-informed *specificity* (**spec**) (Ko et al., 2019); (3) **cosine similarity** between input and response representations, which captures *coherence* (Zhang et al., 2018a).

We follow standards from prior work for human evaluation (Li et al., 2017; Zhang et al., 2018a,b; Xu et al., 2018a). We select 250 prompt-response pairs, and asked 5 judges from MechanicalTurk to rate the responses for each prompt. We evaluate whether the responses are **informative** (Ko et al., 2019; Wu et al., 2018; Shao et al., 2017) and **on topic** with the prompt (Shen et al., 2018; Xu et al., 2018b; Xing et al., 2017), on a scale of 1-5. Average scores are reported. In addition, we evaluate **plausibility** by asking judges whether they think the given response sentence without the prompt can reasonably be uttered, following instructions from Kruszewski et al. (2016). The percentage of plausible ratings are reported.

[4] Although Tao et al. (2018) proposed an unspervised metric, their code is not available.

5.2 Experiment setup

Data We use two datasets in this work: (1) *OpenSubtitles* (Tiedemann, 2009), a collection of movie subtitles widely used in open-domain dialogue generation. We sample 4,173,678 pairs for training and 5,000 pairs for testing from the movie subtitles dataset. Following Li et al. (2017), we remove all pairs with responses shorter than 5 words to improve the quality of the generated responses. (2) *PersonaChat* (Zhang et al., 2018c), a chit-chat dataset collected via crowdsourcing. This is a multi-turn dataset, but we only consider single turn generation in this work. We don't use the personas and false candidate replies. There are 122,458 prompt-response pairs for training and 14,602 pairs for testing. For validation, for reasons described in Section 5.1, we opt for human evaluation of overall response quality on a validation set of 60 prompt-response pairs from PersonaChat.

Settings We use LSTMs with hidden layers of size 500, Adam optimizer (Kingma and Ba, 2015) with learning rate 0.001, $\beta_1 = 0.9, \beta_2 = 0.999$, dropout rate 0.2 for both training and testing, metric embedding dimension 300 and 5 training epochs. We train randomly initialized word embeddings of size 500 for the dialog model and use 300 dimentional GloVe (Pennington et al., 2014) embeddings for reranking classifiers. We generate 15 candidates for reranking per input sentence. To train the 4 reranking classifiers, we use 375,996 positive sentences on Opensubtitles and 110,221 on PersonaChat. We generate one negative sentence per word or phrase in the positive sentences.

Since specificity is the focus of this study, during testing, we use the embedding of the highest specificity level (5) for NIWF and the linguistically informed specificity predictor. For PPW, we observe that the perplexity of generated sentences does not increase beyond the median level (3) during development, hence we use the median level. For comparison, we also report results when all metric levels are set to be the median (level 3).

		human evaluation			automatic metrics			
	Model	Informative	On topic	Plausible	Dist-1	Dist-2	Spec.	Cos.Sim
Opensubtitles	Seq2seq	3.10	3.18	82.4	0.0349	0.138	0.133	0.638
	+coherence	2.59	3.25	90.0	**0.0538***	**0.217***	0.148	0.728*
	+specificity	3.29	**3.40***	72.4	0.0422	0.191*	**0.351***	**0.740***
	+plausibility	**3.53***	**3.39***	76.0	0.0524*	**0.217***	0.342*	0.711*
	MMI-Anti	3.29	3.07	**90.4**	0.0477	0.184	0.136	0.583
	Zhang	2.30	2.34	48.8	0.0305	0.161	0.163	0.648
PersonaChat	Seq2seq	3.01	2.84	86.9	0.0035	0.0126	0.381	0.703
	+coherence	2.71	3.16	**90.0**	0.0097	0.0428	0.279	**0.734***
	+specificity	3.31*	2.92	65.2	0.0135	0.0458	**0.491***	0.669
	+plausibility	**3.39***	**3.45***	71.7	**0.0184***	0.0679	0.483*	0.650
	MMI-Anti	2.79	3.23	79.0	0.0071	0.0289	0.200	0.706
	Zhang	3.07	3.06	44.9	0.0134	**0.0998**	0.400	0.662

Table 2: Results for the base SEQ2SEQ model and each component. (*) denotes significant improvement ($p < 0.05$ with paired bootstrap resampling) over our three baselines and benchmarks (*Seq2seq, MMI-Anti, Zhang*) according to paired bootstrap resampling. Our full reranking model performed the best in informativeness and topic relevance. A drop in plausibility is unavoidable as responses become more specific; the reranking model mitigates this.

Prompt (PersonaChat) thank you . also , the food is amazing . fries are the best thing ever . **Seq2seq** yes it is . do you have a favorite color ? mine is blue . **Ours(Spec)** yes , i am a vegetarian so i eat a lot of meat and bread puffs . **Ours(Spec+Plaus)** i love italian food , but i am a vegan so i eat a lot of curries **MMI-Anti** i agree . i am a vegetarian . **Zhang** i agree . i love to eat and eat .
Prompt (PersonaChat) i work as kindergarten teacher at a school so pretty busy **Seq2seq** what do you do for a living ? i work in a factory **Ours(Spec)** i work at a grocery store and i work for a pharmaceutical company **Ours(Spec+Plaus)** i work as a medical file clerk at a local hospital . **MMI-Anti** what do you do for work ? **Zhang** that is cool . i am a teacher and i love to be able to write .
Prompt (Opensubtitles) The countryside in spring Appears to the lovers in a new light . The air is already warm . **Seq2seq** The snow is in the air . **Ours(Spec)** It is the night of the dawn of the sunset . **Ours(Spec+Plaus)** The light is rising from the skies , and the sun is shining . **MMI-Anti** The light of the sun is rising . **Zhang** The weather in the sky is a , .

Table 3: Example responses.

5.3 Results

We will first discuss results for the overall architecture, then dive into specificity and plausibility.

Overall architecture We evaluate our model against the base SEQ2SEQ for each component: coherence, specificity embeddings, and plausibility reranking (using the mean of all four classifiers). We also benchmark with the MMI-Anti model using mutual information (Li et al., 2016a), as well as Zhang et al. (2018b)'s model that incorporates a Gaussian kernel layer to control for specificity. We ran Zhang's code on our data and set $s = 1$ for PersonaChat and $s = 0.8$ for Opensubtitles when testing.[5] Significance tests are done via Paired Bootstrap Resampling (Berg-Kirkpatrick et al., 2012).

Table 2 shows that for both datasets, our full model with plausibility reranking (according to average posterior of the four classifiers) generates the most informative, relevant and plausible responses. Examples from our full model and the baselines are shown in Table 3.

Incorporating specificity led to more interesting responses, with 6-10% improvement in informativeness and 3-7% improvement in topic relevance. Since the system is trained without any semantics or common sense knowledge, this led to a drop in semantic plausibility. Plausibility reranking successfully mitigates this issue by improving plausibility by 3.6-6.5%. Although responses from MMI-Anti tend to be more plausible than directly using specificity, these responses are not useful if they are even less informative or relevant than the SEQ2SEQ baseline. Zhang et al. (2018b)'s model performed reasonably on PersonaChat but failed on OpenSubtitles.[6] One reason may be that OpenSubtitles is much more diverse in terms of topic and vocabulary, which makes their approach of estimating specificity independent of dialogue

[5] We observed that a higher s on Opensubtitles will result in many grammatical errors.

[6] Their original evaluation was on Chinese Weibo data.

Model	Informative		On topic	
Ours(Spec)	3.31		2.92	
-Linguistic	**3.11***	(-0.20)	**3.10***	(+0.18)
-NIWF	3.23	(-0.08)	3.20*	(+0.28)
-PPW	3.19*	(-0.12)	3.35*	(+0.43)

Table 4: Effect of excluding each specificity metric on PersonaChat. Delta against *Ours(Spec)* are included in parenthesis and (*) denotes significant delta ($p < 0.05$). Excluding linguistically informed specificity led to the greatest drop in informativeness and the slightest increase in topic relevance.

Model	Reranking	Inform.	Topic	Plaus.
Ours(Spec)	—	3.31	2.92	65.2
Ours(Spec +Plausibility)	1-classifier	3.26	3.36*	68.0*
	Max	**3.58***	3.35*	70.0*
	Mean	3.39	**3.45***	**71.7***
	+CoLA	3.45*	3.22*	68.5*
	CoLA	3.36	3.20*	58.0

Table 5: Comparison of different reranking methods on PersonaChat: training a single classifier, using max/mean posterior from four classifiers, and using CoLA. (*) denotes significant improvement over *Ours(Spec)* ($p < 0.05$). Learning multiple classifiers from synthetic data is the most effective.

context less effective. Indeed, we observe unstable word specificity learned across different training rounds and notable grammatical issues on OpenSubtitles. On the contrary, our joint approach gave stable performance on both datasets.

On PersonaChat, our coherence component led to improvements in topic relevance and cosine similarity, while specificity improved topic relevance and diversity, which is an intuitive result. On OpenSubtitles, coherence led to increased diversity while specificity led to a decrease. We looked into this and found that length trade-off is at play since the Distinct measures normalize by length of all generated responses: coherence led to diverse but short responses while specificity increased length. On human evaluation, they complement each other and using both gave better overall results. While reranking clearly did improve plausibility, there is also notable improvement in informativeness. This shows that informativeness is not only a frequency-only issue, or even a specificity-only issue, and that semantic plausibility plays an important role. Since the automatic metrics do not capture plausibility information in the sentence, it is unsurprising that they did not improve with plausibility added in.

We also study the effect of maxing out specificity and coherence levels at test time vs. using an uninformative level (median). Using median significantly improved informativeness and diversity (distinct-2) on PersonaChat by 0.90 and 0.53, and did not improve topic relevance. Similar but insignificant improvements are observed on OpenSubtitles. On the other hand, using the maximum levels led to significant improvements over the baseline or the median level on all metrics.

Specificity We now dive into a more detailed analysis for each specificity metric on PersonaChat. Table 4 shows human evaluation of informativeness, topic relevance and plausibility for the non-reranking model minus one specificity metric. Notably, excluding the linguistic based metric resulted in the largest drop in informativeness and relevance. Frequency based NIWF has the least impact on informativeness, indicating that specificity in dialogue is a multi-faceted issue and that the linguistically-informed notion is the most suitable. If none of the specificity metrics are included, topic relevance scores improve. This is because increasing specificity leads to fewer generic responses, yet they are more likely to be judged "on topic" by humans.

Plausibility We compare several different settings for plausibility reranking. Table 5 shows three ways of using the synthetically generated sentences discussed in Section 4: (1) **1-classifier**, which trains one classifier to distinguish true responses vs. all generated ones; (2) **Max**, which trains separate classifiers and take the maximum posterior probability (recall that higher posterior means less plausible responses); (3) **Mean**, which trains separate classifiers and averages the posterior probability. For all classifiers, at least 72% of the responses ranked top 50% on a balanced test set are true responses.

All three reranking methods helped, however, using one classifier is less effective than training and aggregating separate classifiers for each type of semantic implausibility. The latter not only improved plausibility but also informativeness and topic relevance. Using Max vs Mean yields comparable results in terms of plausibility, although Max improves informativeness more while Mean improves topic relevance more.

We also experimented with training an additional classifier (of the same architecture) on

Seq2seq	Ours(Spec)	Ours(Spec +Plausibility)	MMI -Anti	Zhang
82.4	78.7	82.6	90.0	61.5

Table 6: Percentage of sentences judged grammatical on OpenSubtitles.

the Corpus of Linguistic Acceptability (Warstadt et al., 2018), a dataset consisting of linguistically acceptable vs. unacceptable sentences. However, looking at results from PersonaChat, reranking using CoLA did not improve plausibility although is of slight help for informativeness and topic relevance. Combining CoLA with the other four classifiers decreased plausibility.

Grammaticality Finally, since the function word substitution aspect of our synthetic sentences is related to grammar, we also conduct human evaluation of grammaticality on OpenSubtitles. We did not evaluate on PersonaChat because almost all generate responses of our model we inspected are grammatically correct. Here annotators are asked to judge whether a sentence is grammatical vs. not. Results are shown in Table 6.

Informative and interesting responses that are the result of increasing specificity also made the model more prone to grammatical errors, but adding reranking completely mitigated this issue and grammaticality results are the same as the base model that generates much shorter, canned universal responses. MMI gave the best grammaticality; however, these response are not useful if they are even less informative or relevant than the SEQ2SEQ baseline. Zhang et al. (2018b)'s model generated more complicated sentences, but has worse grammar. Again we suspect that this is because of the lack of interaction between specificity estimates and dialogue context in their model.

6 Conclusion

We presented a new method to incorporate specificity information and semantic plausibility in SEQ2SEQ models. We showed that apart from frequency-based specificity metrics explored in prior work, information-theoretic and linguistically informed specificity improve the specificity of the responses. We proposed a reranking method aimed at improving the semantic plausibility of specific responses. Results showed that our method improved human ratings on informativeness, plausibility and grammaticality on both open domain and chit-chat datasets.

Acknowledgments

This work was partially supported by the NSF Grant IIS-1850153, and an Amazon Alexa Graduate Fellowship. We thank the anonymous reviewers for their helpful feedback.

References

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating more interesting responses in neural conversation models with distributional constraints. In *EMNLP*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *EMNLP*.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Jennifer C. Lai, and Robert L. Mercer. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*.

Peter Dixon. 1987. The processing of organizational and component step information in written directions. *Journal of memory and language*, 26(1):24–35.

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. Frage: Frequency-agnostic word representation. In *NIPS*.

Joshua Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15:403–434.

A.L. Gorin, J.H. Wright, G. Riccardi, A. Abella, and T. Alonso. 2000. Semantic information processing of spoken language. In *ATR Workshop on Multi-Lingual Speech Communication*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *ACL*.

Chenyang Huang, Osmar R. ZaIane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *NAACL*.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Domain agnostic real-valued specificity prediction. In *AAAI*.

Germán Kruszewski, Denis Paperno, Raffaella Bernardi, and Marco Baroni. 2016. There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Computational Linguistics*, 42(4):637–660.

Karla A Lassonde and Edward J O'Brien. 2009. Contextual specificity in the activation of predictive inferences. *Discourse Processes*, 46(5):426–438.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *ACL*.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016c. A simple, fast diverse decoding algorithm for neural generation. In *arXiv CS.CL*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016d. Deep reinforcement learning for dialogue generation. In *EMNLP*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *EMNLP*.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *AAAI*.

Junyi Jessy Li and Ani Nenkova. 2016. The instantiation discourse relation: A corpus analysis of its properties and improved detection. In *NAACL*.

Junyi Jessy Li, Bridget O'Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. 2016e. Improving the annotation of sentence specificity. In *LREC*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

Pierre Lison and Serge Bibauw. 2017. Not all dialogues are created equal: Instance weighting for neural conversational models. In *SIGDIAL*.

Chia-Wei Liu, Ryan Lowe, Iulian V. 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 *EMNLP*.

Yahui Liu, Victoria, Jun Gao, Xiaojiang Liu, Jian Yao, and Shuming Shi. 2018. Towards less generic responses in neural conversation models:a statistical re-weighting method. In *EMNLP*.

Annie Louis and Ani Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *IJCNLP*.

Luca Lugini and Diane Litman. 2017. Predicting specificity in classroom discussion. In *Workshop on Innovative Use of NLP for Building Educational Applications*.

Tomáš Mikolov, Stefan Kombrink, Anoop Deoras, Lukáš Burget, and Jan Černockỳ. 2011. RNNLM - recurrent neural network language modeling toolkit. In *ASRU 2011 Demo Session*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *EMNLP*.

Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018. Improving variational encoder-decoders in dialogue generation. In *AAAI*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *NAACL*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI*.

Jörg Tiedemann. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In *RANLP*.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. In *NAACL*.

Alex Warstadt, Amanpreet Singh, and Sam Bowman. 2018. Neural network acceptability judgments. In *arXiv CS.CL*.

Yu Wu, Wei Wu, Dejian Yang, Can Xu, Zhoujun Li, and Ming Zhou. 2018. Neural response generation with dynamic vocabularies. In *AAAI*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *AAAI*.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018a. DP-GAN: Diversity-promoting generative adversarial network for generating informative and diversified text. In *EMNLP*.

Xinnuo Xu, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. 2018b. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *EMNLP*.

Kaisheng Yao, Baolin Peng, Geoffrey Zweig, and Kam-Fai Wong. 2016. An attentional neural conversation model with improved specificity. In *arXiv CS.CL*.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018a. Reinforcing coherence for sequence to sequence model in dialogue generation. In *IJCAI*.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, YanyanLan, Jun Xu, and Xueqi Cheng. 2018b. Learning to control the specificity in neural response generation. In *ACL*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018c. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*.

Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In *AAAI*.