Deexaggeration

Li Kong[1] , **Chuanyi Li**[1] and **Vincent Ng**[2]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2]Human Language Technology Research Institute, University of Texas at Dallas, Richardson, Texas, USA
kl_nju@126.com, lcy@nju.edu.cn, vince@hlt.utdallas.edu

Abstract

We introduce a new task in hyperbole processing, deexaggeration, which concerns the recovery of the meaning of what is being exaggerated in a hyperbolic sentence in the form of a structured representation. In this paper, we lay the groundwork for the computational study of understanding hyperbole by (1) defining a structured representation to encode what is being exaggerated in a hyperbole in a non-hyperbolic manner, (2) annotating the hyperbolic sentences in two existing datasets, HYPO and HYPO-cn, using this structured representation, (3) conducting an empirical analysis of our annotated corpora, and (4) presenting preliminary results on the deexaggeration task.

1 Introduction

Recent years have seen a surge of interest in the automatic processing of figurative language in the natural language processing (NLP) community. Much of the work on figurative language processing has focused on metaphor and metonymy [Tsvetkov *et al.*, 2014], and more recently, sarcasm [Hazarika *et al.*, 2018], idioms [Liu and Hwa, 2018], and puns [He *et al.*, 2019]. In particular, hyperbole, also known as exaggeration, is a relatively under-studied phenomenon in the community. This is somewhat surprising, especially given that the prevalence of hyperbole as a rhetorical device is only second to metaphor [Kreuz *et al.*, 1996]. Humans exaggerate in different situations for various purposes, such as creating amusement, expressing emotion and drawing attention [Li, 2013].

The computational treatment of hyperbole is still in its infancy. So far, automatic hyperbole processing has focused on automatic hyperbole detection (i.e., determining whether a sentence is a hyperbole or not). To facilitate the study of automatic hyperbole detection, Troiano et al. [2018] and Kong et al. [2020] have assembled HYPO and HYPO-cn respectively, which are English and Chinese datasets composed of a set of hyperbolic sentences and their non-hyperbolic counterparts. Research in the broader figurative language processing community has exhibited a similar trend, focusing almost exclusively on *detecting* whether a certain rhetorical device is used in a sentence and *extracting* the words/phrases in a sentence in which a certain rhetorical device is used.

Input document: Tom felt that the team's loss in yesterday's game was the biggest disaster of the 21st century.

Question: Which of the following can be inferred?
A. A big disaster happened yesterday.
B. Tom felt sad.

Table 1: A reading comprehension example.

To benefit high-level NLP applications, however, figurative language processing research cannot merely perform detection and extraction. Consider, for instance, machine comprehension, which has increasingly focused on answering questions that require inference as well as a deep understanding of the input text. An example question is shown in Table 1, where choosing the correct answer (option B) requires that the hyperbolic sentence in the input document be interpreted properly. As the vast majority of existing NLP systems, including many high-level NLP applications such as text summarizers and machine translation (MT) systems, are trained on text documents that can be interpreted literally, it is likely that they will fail to properly understand texts with a non-literal meaning. Consequently, being able to *interpret* figurative language in general and hyperboles in particular will stand to expand the capabilities of today's NLP applications.

In light of the above discussion, our work on hyperbole processing in this paper focuses on hyperbole *interpretation*. More specifically, interpreting a hyperbolic sentence involves determining its literal meaning. Our focus on interpretation sets our work apart from essentially all existing work on figurative language processing, which has focused on *detection* and *extraction*, as mentioned above. In particular, to our knowledge, no work on figurative language processing has focused on the challenging task of automatically recovering the literal meaning of a non-literal sentence.

A simple way to interpret hyperbolic sentences would be to recast the task as MT, where a hyperbolic sentence is "translated" into its non-hyperbolic counterpart so that the resulting sentence can be interpreted in a literal manner using existing language processing tools. We experimented with this idea by training a seq2seq model to translate hyperbolic sentences into non-hyperbolic sentences, but the translation quality is poor. An examination of the results reveals that the small size of the parallel corpora provided by HYPO and HYPO-cn, which have 709 and 2680 parallel sentences respectively, is the primary cause of failure.

Motivated by this observation, we propose a different approach that draws inspirations from work in opinion mining, where the opinion of a sentence is represented as a tuple (g, s, h, t) [Liu, 2015] with g being the target of the opinion (e.g., the topic, event, object, etc., towards which the sentiment is expressed), s being the sentiment expressed, h being the holder of the opinion, and t being the time when the opinion is expressed. We hypothesize that a tuple can be similarly defined to represent what is being exaggerated in a hyperbole in a non-hyperbolic manner. Adopting a structured representation like this potentially allows us to extract/learn each element in the tuple independently, thus reducing the complexity of the learning task. Just as the opinion representation described above is intended to encode the opinion expressed in a sentence, our exaggeration representation is intended to encode the exaggeration expressed in a hyperbolic sentence in a non-hyperbolic manner. In particular, it is not intended to encode the meaning of the given hyperbolic sentence in a non-hyperbolic manner, which is what MT approach would produce. To exemplify, consider the following sentence:

[1] The driver disappeared after hitting someone.

What is exaggerated in [1] is the use of the word "disappeared" to describe the speed with which the driver left the scene after the accident. This is what our representation is intended to focus on, not "after hitting someone". We will henceforth refer to the task of producing a structured representation of what is exaggerated in a hyperbolic sentence in a non-hyperbolic manner as *deexaggeration*.

Our goal in this paper is to lay the groundwork for the computational study of deexaggeration. First, we define structured representations to encode what is being exaggerated in hyperbolic sentences in a non-hyperbolic manner. Second, we annotate the hyperbolic sentences in two corpora using the meaning representation we define, as progress on automatic hyperbole processing is hindered in part by the lack of annotated corpora. While we could assemble and annotate our own set of hyperbolic sentences, we chose to annotate the sentences in HYPO and HYPO-cn because (1) the hyperbolic sentences in these corpora have been carefully verified to be hyperbolic, and (2) we see advantages in augmenting existing corpora with additional linguistic annotations, as the annotations can be consolidated to train complex models that allow multiple related tasks to be jointly learned. Third, we conduct an empirical analysis of our annotations. Finally, we train models to obtain preliminary results on the deexaggeration task, which can serve as baseline results for future work.

2 Related Work

2.1 Linguistic Studies on Hyperbole

Hyperbole has long been studied in pragmatic linguistics. Mora [2009] studies the production of hyperboles from a semantic perspective by constructing a taxonomy in which hyperboles are categorized along two dimensions, quantitative (inflating a quantitative or objective property) and qualitative (inflating a qualitative or subjective property). Zhao and Lu [2013] argue that a good hyperbolic sentence should exaggerate the reality and yet is logical. McCarthy and Carter [2004] emphasize the interactive nature of hyperbole, as listener reaction is important for the interpretation of a hyperbole.

A lot of work has focused on manually analyzing how humans exaggerate. Examining the novel "Er Ma", Liao and Ge [2014] conclude that hyperboles can be expressed via (1) an upsurge on a semantic scale, which can be qualitative or quantitative, or (2) other rhetorical devices, including personification and metaphor. Studying Mo Yan's novel "Sandalwood Punishment", Zhang [2016] points out that exaggeration may involve (1) an upsurge on a semantic scale or (2) presenting two events out of their typical temporal order, and concludes that exaggeration can be expressed using one of eight strategies: Direct Hyperbole (which occurs when other rhetorics are not involved), Extreme Quantity (semantic upsurge on a quantitative scale), Extreme Quality (semantic upsurge on a qualitative scale), Double Negation, Metaphor, Personification, Comparison, and Other.

2.2 Figurative Language Comprehension

Recent years have seen growing interest in figurative language comprehension. For metaphor processing, Su *et al.* [2020] propose a hierarchical semantic model to perform metaphor comprehension considering cultural factors; Rivera *et al.* [2020] build a neural network to detect the metaphoricity of adjective-noun pairs using pre-trained word embeddings and word similarity; Zhang *et al.* [2019] use an attention network based on subject–predicate and verb–object relations to identify Chinese verb metaphors; and Chen *et al.* [2019] detect Chinese metaphors using various kinds of cultural background information such as radicals representing body parts, instruments, materials, and movements. For sarcasm detection, Sundararajan and Palanisamy [2020] identify the level of hurt or the true intent behind sarcastic text and propose a rule-based approach to determine the type of sarcasm; and Hazarika *et al.* [2018] extract contextual information together with user embeddings in online social media discussions. For idiom analysis, Liu and Hwa [2018] identify the intended usage of an idiom by treating possible usages as a latent variable in probabilistic models and training them in a linguistically motivated feature space; Liu *et al.* [2019] highlight the importance of idioms in writing and leverage a neural translation framework to realize idiom recommendation; and Colston and Keller [1998] study how humans comprehend irony in expressing surprise. For homographic pun detection, Diao *et al.* [2019] use a contextualized representation with a gated attention. Finally, for the study of euphemistic and dysphemistic language, Felt and Riloff [2020] extract near-synonym phrases for three topics via bootstrapping and identify such language using lexical sentiment cues and contextual sentiment analysis.

2.3 Figurative Language Generation

Research on figurative language generation has mainly focused on the generation of puns and metaphors.

Early work on pun generation is template-based. For example, Hong and Ong [2009] use phonetic and semantic linguistic resources to extract word relationships in puns and store the knowledge in the form of a template. Valitutti *et al.* [2009] present an interactive system for producing humorous puns obtained through word replacement performed on familiar expressions, where the replacement is selected according to phonetic similarity and semantic constraints expressing semantic opposition or evoking ridiculousness. Yu *et al.* [2018] are the first to propose a seq2seq framework for pun generation, where a conditional neural language model is trained on a general text corpus to generate homographic puns with a specially designed decoding algorithm.

As for metaphor generation, Yu and Wan [2019] extract metaphorically used verbs with their metaphorical senses in an unsupervised manner and train a neural language model to generate metaphors conveying the assigned metaphorical senses. Stowe *et al.* [2021] improve the generation of the metaphoric version of a literal sentence using a controlled generation process that incorporates information based on conceptual metaphor theory.

3 Corpus Annotation

3.1 Corpora

For annotation, we use two corpora previously assembled for research on automatic hyperbole detection, HYPO (English), and HYPO-cn (Chinese).

HYPO is composed of 709 hyperbolic sentences, each of which has a non-hyperbolic version. HYPO-cn is composed of 4762 sentences, of which 2680 are hyperbolic and 2082 are non-hyperbolic. These 4762 sentences can be partitioned into 700 *sets*, where the sentences in each set are hyperbolic/non-hyperbolic versions of each other.[1] As mentioned in the introduction, we used only the hyperbolic sentences in each corpus for annotation, but verified that the tuple we annotated for each hyperbolic sentence encodes the same meaning as its non-hyperbolic version.

3.2 Structured Representation

Next, we design a structured representation that can be used to encode what is being exaggerated in a hyperbolic sentence in a non-hyperbolic manner. According to linguistic studies on hyperbole (e.g., Mora [2009], Stanivukovic [2007]), one can exaggerate either a particular attribute of an *entity* or the manner in which an action took place during an *event*. Below we refer to these two types of exaggeration as *static* exaggeration and *dynamic* exaggeration respectively and design a structured representation for each type of exaggeration.

Static Exaggeration

In static exaggeration, the author's intent is to exaggerate the *description* associated with a particular *aspect* of a *target* entity. In general, the target can be a person, an object, or a location. We borrow from aspect-based sentiment analysis the term *aspect*, which is used to refer to an attribute of a product or service (e.g., the *battery* of a laptop), but overload it so that it can also refer to an object or an abstract concept possessed by or associated with the target. Consider the sentence:

[2] Mary has countless toys.

Here, "countless" is used to exaggerate the number of toys.

To encode what is being exaggerated in a sentence that exhibits static exaggeration in a non-hyperbolic manner, we design a 3-tuple (t, a, v), where t is the target, a is the aspect of t whose description is being exaggerated, and v is the value of a, which can be viewed as the "deexaggerated" equivalent of the exaggerated description. For instance, the tuple that should be produced from [2] is ("Mary", "toys", "lots of") , where "lots of" is the value of the aspect "toys" that is the deexaggerated version of "countless". We require that t and a be nouns (or noun phrases) and v be an adjective.

While the target always appears in a hyperbolic sentence, the aspect and its value may not be lexically realized and should be inferred. Consider the following sentence:

[3] The whole world is against you.

[3] means that you have bad luck, so the tuple to be produced is ("you", "luck", "bad"). While the target "you" appears in the original sentence, the aspect and its value do not.

Dynamic Exaggeration

In dynamic exaggeration, the author's intent is to exaggerate the *manner* in which an *action* that took place during an event is performed by an *agent* on a *patient*. Consider the following sentence:

[4] John would rather die than kill the innocent.

[4] means that "John would not kill the innocent". Here, the action is "kill" and has "John" as its agent and "the innocent" as its patient. The phrase "rather die than" exaggerates the point that John would not kill the innocent *no matter what*.

We design a 4-tuple (a, p, v, m) to encode the meaning of what is being exaggerated in a sentence that exhibits dynamic exaggeration, where m is the deexaggerated equivalent of the exaggerated expression used to describe the manner in which action v in an event is performed, and a and p are respectively the agent and the patient of v. Given [4], the tuple ("John", "the innocent", "kill", "not") should be produced. We require that a and p be nouns, v be a verb, and m be an adverb.

Note that the agent and the patient may not appear in a hyperbolic sentence (e.g., if the sentence is imperative or the arguments associated with the action/event are implicit). In that case, we will set a and p to NULL, meaning that they do not need to be inferred if they do not appear in the original sentence. In contrast, v and m may need to be inferred.

3.3 Annotation Procedure

For annotation, we hired two native speakers of English and two native speakers of Chinese. These annotators are graduate students in NLP (none of them are the authors) and have received a one-hour tutorial on exaggeration in which we (1) presented the language-independent criteria of exaggeration described in Troiano *et al.* [2018] and (2) introduced our structured representations as well as guidelines and examples on how to use them to annotate hyperbolic sentences. After that, the two native speakers of English were asked to independently label each hyperbolic sentence in HYPO with our structured representations, and the two native speakers of Chinese were asked to do the same for the sentences in HYPO-cn.

[1] Although some sentences come from the same set, they were presented to our human annotators as *independent* sentences for the purpose of annotation.

English										
	All	Static				Dynamic				
Agreement criterion	Sentences	All	Target	Aspect	Value	All	Agent	Patient	Action	Manner
Strict	0.72	0.77	0.91	0.84	0.77	0.63	0.93	0.78	0.87	0.80
Same meaning as agreement	0.81	0.85	0.92	0.88	0.85	0.78	0.94	0.85	0.88	0.85
Correct interpretations as agreement	0.88	0.90	0.93	0.92	0.91	0.89	0.94	0.91	0.91	0.89
Chinese										
	All	Static				Dynamic				
Agreement criterion	Sentences	All	Target	Aspect	Value	All	Agent	Patient	Action	Manner
Strict	0.65	0.68	0.90	0.81	0.75	0.58	0.90	0.84	0.80	0.77
Same meaning as agreement	0.80	0.83	0.91	0.88	0.85	0.79	0.90	0.87	0.87	0.85
Correct interpretations as agreement	0.89	0.91	0.93	0.92	0.92	0.89	0.92	0.90	0.92	0.91

Table 2: Inter-annotator agreement on tuples as measured by the Dice Coefficient.

3.4 Inter-annotator Agreement

In this subsection, we report the inter-annotator agreement for two key annotation tasks.

Static/Dynamic Labeling

Given a hyperbolic sentence, the first annotation task involves determining whether static or dynamic exaggeration was used. We reached near-perfect agreement on this task, having Cohen's kappa values of 0.978 for English and 0.985 for Chinese. These numbers suggest the simplicity of the task. Every instance of disagreement was subsequently resolved via discussion.

Tuple Creation

Tuple creation, our second annotation task, is substantially more difficult than the first annotation task. Given its complexity, we design three *agreement criteria* to compute inter-annotator agreement.

Strict. This is the strictest of the three agreement criteria in which two annotations are counted as an agreement if and only if they are identical.

Same meaning as agreement. Our annotation task requires extracting text spans from the hyperbolic sentences and in many cases performing inference on the hyperbolic sentences to generate new words/phrases. Hence, the Strict agreement is arguably overly strict. Our second agreement criterion is a relaxation of the Strict criterion: it is the same as Strict except that two annotations are counted as an agreement if they are synonyms or the tuples in which they appear have the same meaning. For instance, two Target annotations “water in the sea” and “seawater” are counted as an agreement w.r.t. this second criterion, and so are the Aspect annotations ”weight” and ”body weight”. This agreement criterion has a greater impact on those elements of the tuples whose values often have to be inferred (e.g., Aspect, Value, Manner) than those whose values can often be extracted from the hyperbolic sentences (e.g., Target, Agent).

Correct interpretations as agreement. While the second agreement criterion is more realistic for our annotation task, we believe it does not go far enough. In particular, some hyperbolic sentences may leave open more than one interpretation of what is being exaggerated. Consider the sentence “The admirers of Mary can fill a train”. One annotator annotated the sentence with the tuple (“admirers”, “number”, “large”), while another annotated it as (“Mary”, “popularity”, “high”). In other words, one thought that the number of admirers was exaggerated, whereas the other focused on Mary's popularity. We thus design a third agreement criterion, which is a relaxation of the second one. Specifically, this criterion is the same as the second one except that two annotations are counted as an agreement if they are different but are both correct interpretations of the sentence.

While computing agreement using the Strict criterion can be automated, computing agreement using other two cannot.[2] Therefore, we employ two additional graduate NLP students (one for HYPO and one for HYPO-cn) who are experts in computational exaggeration and did not participate in the annotation process described in Section 3.3 to determine whether two annotations agree.

Table 2 shows the results of inter-annotator agreement, where *agreement* is reported in terms of the *Dice Coefficient* according to each of the three agreement criteria. We report agreement at different levels. In the “All Sentences”, “Static All”, and “Dynamic All” columns, we report agreement at the tuple level for all sentences, all “Static” sentences, and all “Dynamic” sentences, respectively. For instance, the 0.72 agreement value for English in the “All Sentences” column w.r.t. the Strict criterion means that 28% of the tuples produced by the two annotators on all the hyperbolic sentences in the English dataset are not identical to each other. As can be seen, we also report agreement w.r.t. each element.

Perhaps not surprisingly, the agreement score increases as we relax the agreement criterion. In addition, the elements that have lower agreement rates are those that need to be inferred more frequently, such as Aspect, Value, Action, and Manner. Nevertheless, using the most relaxed agreement criterion, the agreement scores are above 0.88 for all of the elements. All of the remaining disagreements can be attributed to misinterpretation of the meaning of the given hyperbolic sentence or what is being exaggerated by one of the annotators. For instance, given the sentence ”Stop talking! My ears are calloused!”, the correct tuple should be (”you”, ”talking”, ”excessive”), while one annotator incorrectly annotated the sentence as (”my ears”, ”comfort level”, ”low”). In the end, all the incorrect tuples are discarded, and all the tuples that

[2] Computing disagreement using the second criterion could be automated using a lexical paraphraser, but computing disagreement using the third criterion would be difficult to automate.

	English		Chinese	
	Static	Dynamic	Static	Dynamic
# sentences	417	292	1741	939
# tuples	447	317	1858	1026

Table 3: Statistics on our annotations.

Static			Dynamic		
Element	EN	ZH	Element	EN	ZH
Target	100	100	Agent	99.5	99
Aspect	16.7	39.9	Action	37.9	62.3
Value	8.7	24.3	Patient	92.6	92.3
			Manner	59.2	53.6

Table 4: Percentage of the values of each element that appear in the corresponding hyperbolic sentence for English (EN) and Chinese (ZH).

are not considered as a disagreement by the second and third agreement criteria are retained.

3.5 Annotation Statistics

Statistics on our annotations are shown in Table 3. 65% of the Chinese sentences are labeled as static, while the corresponding number for English is 58.8%. Also, the number of tuples exceeds the number of hyperbolic sentences since each sentence may involve more than one correct tuple.

Table 4 shows the percentage of the values of each element in the tuple that also appear in the corresponding hyperbolic sentence. As can be seen, elements like Target, Agent, and Patient can usually be found in the hyperbolic sentences, whereas the others frequently need to be inferred.

4 Annotation Analysis

4.1 Annotation Difficulty Levels

Our annotators reported that some sentences are easier to annotate than others. In particular, a hyperbolic sentence is easier to annotate if all elements of the tuple can be directly extracted from it than if many elements need to be inferred. Guided by this intuition, they design the following *annotation difficulty levels* in an attempt to characterize how difficult it is to annotate a given hyperbolic sentence.

For static exaggeration, the annotators came up with four difficulty levels. In Level 1 (Easy), all three elements can be extracted from the sentence. In Level 2 (Fair), Value needs to be inferred, but the answer is generally obvious. In Level 3 (Difficult), Value also needs to be inferred, but the inference is harder as more than one interpretation is plausible. Finally, in Level 4 (Challenging), it is hard to infer any element.

For dynamic exaggeration, they came up with three difficulty levels. In Level 1 (Easy), Action and Manner can be extracted or easily inferred from the sentence. In Level 2 (Fair), Action can be extracted or inferred, but Manner is absent and needs to be inferred. Finally, in Level 3 (Difficult), Action is absent in the sentence and needs to be inferred.

Table 5 shows the percentage distribution of the sentences over the difficulty levels. Overall, the English sentences expressing both static and dynamic exaggeration are perceived to be harder to annotate than their Chinese counterparts.

Static			Dynamic		
Level	EN	ZH	Level	EN	ZH
1	43.6	50.9	1	48.0	53.4
2	33.1	33.5	2	29.8	31.9
3	15.8	10.1	3	22.2	14.7
4	7.5	5.5			

Table 5: Distribution of hyperbolic sentences over the difficulty levels for English (EN) and Chinese (ZH).

4.2 Relation to Exaggeration Strategies

Each hyperbolic sentence in HYPO-cn is annotated with one of 11 exaggeration strategies[3], which are strategies humans commonly employ to produce hyperboles. Table 6 provides an explanation. An interesting question is: is there any correlation between how difficult it is to annotate a hyperbolic sentence with our tuple and the exaggeration strategy?

Table 7 shows for each strategy the distribution of its sentences over the difficulty levels. We found several interesting correlations between strategies and difficulty levels. First, sentences that employ Quantity Concepts, Extreme Cases, and Comparison as their strategies are generally easy to annotate. For instance, Comparison sentences tend to compare a target's aspect with another person/object's value, so Target, Aspect, and Value are explicitly mentioned. Second, sentences that employ Common Sayings, Supernatural Concepts, Human Body, and Fictitious Scene as their strategies are generally hard to annotate. For instance, when supernatural concepts are used, many elements in the tuple are omitted (e.g., "She is a fairy"), thus increasing inference load on the reader. Finally, sentences that employ Rhetorics, About Life, and About Nature have a less skewed distribution over the difficulty levels. Generally speaking, however, our analysis reveals that sentences employing these strategies do not have obvious regularities in terms of syntactic structure and the way meaning is conveyed.

5 Evaluation

In this section, we present preliminary results on the task of predicting the tuple associated with a hyperbolic sentence, with the goal of gauging the difficulty of the task. We decompose the task into two subtasks. First, given a hyperbolic sentence, we classify it as an instance of static or dynamic exaggeration. Second, given a hyperbolic sentence and its *gold* static/dynamic class label, we predict the elements of the tuple. Note that we assume gold static/dynamic classes as input to our second task because our goal is to get a better idea of how well we can predict the tuple without being adversely affected by the noise inherited from the first task.

5.1 Approaches

For the first task (static vs. dynamic prediction), we implement a binary classifier with a pre-trained BERT model (albert_tiny[4] for Chinese and bert_small[5] for English). We fine-

[3] HYPO does not contain such annotations, so we manually label the sentences in HYPO with these strategies.

[4] https://github.com/brightmart/albert_zh

[5] https://github.com/google-research/bert

Strategy	Explanation	Example
Quantity concepts	use numbers to overstate	The buyer needs a hundred eyes.
Extreme cases	includes non-exceptionality, non-existence, etc.	It is the end of the world.
Common sayings	idioms, proverbs, etc.	The style is the man.
Rhetorics	use of other rhetorical devices in hyperboles	Technology annihilates humanity.
Comparison	use a reference to highlight/exaggerate	Soon I'll be famous beyond belief.
Supernatural concepts	reference to prophets, gods, immortals	You look as white as a ghost.
About life	includes the concept of bringing/destroying life	I would give my life for a coffee.
State of human body	describes an unusual state of the body	My heart is bleeding right now.
About nature	reference to entities in nature	Love you to the moon and back.
Fictitious scene	use of an imaginary scene to overstate	If you cough in Siberia, she'll hear you.
Impossible ordering	describes an out-of-order event sequence	He has been teaching since Stone Age.

Table 6: An overview of the 11 exaggeration strategies.

	English							Chinese						
	Static				Dynamic			Static				Dynamic		
Strategy	1	2	3	4	1	2	3	1	2	3	4	1	2	3
Quantity concepts	60.9	30.4	8.7	0.0	76.9	15.4	7.7	51.7	36.6	10.5	1.2	81.5	18.5	0.0
Extreme cases	75.9	13.9	6.3	3.8	74.5	19.1	6.4	59.9	28.9	7.9	3.3	73.3	18.1	8.6
Common sayings	44.4	33.3	11.1	11.1	50.0	16.7	33.3	50.2	31.3	10.9	7.5	52.7	31.9	15.4
Rhetorics	19.4	50.7	21.5	8.3	26.5	38.2	35.3	45.7	41.4	8.6	4.3	30.9	46.3	22.8
Comparison	59.5	31.0	7.1	2.4	100	0.0	0.0	81.9	12.7	3.9	1.5	60.0	36.0	4.0
Supernatural concepts	35.7	35.7	21.4	7.1	50.0	21.4	28.6	39.7	43.1	6.9	10.3	41.3	32.6	26.1
About life	53.3	26.7	13.3	6.7	52.9	29.4	17.6	48.8	29.3	12.2	9.8	55.2	20.7	24.1
Human body	35.7	14.3	28.6	21.4	28.6	28.6	42.9	59.6	24.2	7.5	8.7	53.6	27.3	19.1
About nature	39.1	30.4	17.4	13.0	33.3	57.1	9.5	33.0	40.2	19.6	7.1	39.0	54.2	6.8
Fictitious scene	47.5	20.0	20.0	12.5	30.0	36.7	33.3	32.2	43.3	15.9	8.6	44.9	32.7	22.4
Impossible ordering	0.0	0.0	0.0	0.0	0.0	100	0.0	33.3	66.7	0.0	0.0	25.0	75.0	0.0

Table 7: Distribution of the sentences for each exaggeration strategy over difficulty levels.

tune the BERT layers together with the following layers on our training data. A [CLS] symbol is inserted at the beginning of the input sentence, whose output embedding is used as the semantic representation of the input (hyperbolic) sentence to classify the sentence.

For the second task (tuple prediction), to seek preliminary results, we employ three approaches.

Our first approach is a *generation* approach. For each element ele, we train a BERT-based seq2seq model, which is an implementation of Microsoft's UniLM [Dong *et al.*, 2019], as follows. We first concatenate the hyperbolic sentence and the value to be predicted as the input and the output of the model, and then train the model to predict the value by masking the hyperbolic sentence in the output. Since we have seven elements, we train seven seq2seq models for each language.

Our second approach is an *extraction + classification* approach. For each element ele, we train a sequence tagger, specifically a BERT-BiLSTM-CRF model, to extract the value of ele directly from the input sentence. The output embedding of (English) words / (Chinese) characters of the pretrained BERT model are treated as the input of a bidirectional layer, and the final CRF layer will generate a probability distribution of the possible tags. Since we adopt the IOB convention, there are three possible tags. As noted before, there are four elements whose values are less likely to appear in the given hyperbolic sentence, so this extraction-based approach may not work well. Consequently, we additionally employ a classification-based approach, where we train one classifier for each of these elements to predict its value. This classifier is implemented in the same way as the one used for the static/dynamic prediction task. The possible class values of the classifier for element ele are the set of values of ele that have appeared in the training set.

Our third approach is also an *extraction + classification* approach but differs from the second one in that (1) it *jointly* trains the sequence tagger and the classifier and (2) the different elements of a tuple are predicted using the *same* model. Specifically, the tagger predicts Target (if the tuple is "static") or Agent and Patient (if the tuple is "dynamic" tuple), whereas the classifier predicts Aspect and Value (if the tuple is "static") or Action and Manner (if the tuple is "dynamic"). To do so, we add a classification layer to the BERT-BiLSTM-CRF model used in the second approach. The classifier's input is created by performing a *reduce_max* to extract the maximal probability value from each position in the BiLSTM hidden layer output. Note that eventually four models are trained for handling static/dynamic English/Chinese hyperbolic sentences.

5.2 Experimental Setup

All results are obtained via five-fold cross-validation experiments.[6] In each fold experiment, we use three folds for training, one fold for development, and one fold for testing.

[6] Recall from Section 3.1 that the hyperbolic sentences in HYPO-cn can be divided into 700 sets. In our partition of the Chinese hyperboles into folds, we ensure that the sentences from the same set also appear in the same fold.

Language	Element	First Approach			Second Approach						Third Approach		
		Seq2Seq			CRF			Classifier			Joint CRF-Classifier		
		R	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
English	Target	48.9	51.3	49.2	48.2	52.8	50.4	–	–	–	57.8	62.7	**59.1**
	Aspect	15.9	12.7	13.0	10.3	9.8	10.0	12.2	13.4	12.8	16.9	20.8	**17.9**
	Value	16.3	13.3	13.8	6.5	7.6	7.2	12.6	15.2	13.8	17.5	19.6	**17.8**
	Agent	53.5	58.9	55.4	57.9	64.9	61.0	–	–	–	63.3	66.7	**64.3**
	Patient	32.8	34.6	32.6	32.9	27.0	29.7	–	–	–	32.8	36.3	**33.6**
	Action	12.8	13.3	13.6	18.7	23.4	**21.2**	15.5	13.8	15.3	19.7	22.5	20.6
	Manner	12.9	11.7	11.6	20.4	14.6	18.2	24.1	16.9	**20.2**	16.1	18.8	15.8
Chinese	Target	41.6	36.0	35.8	68.4	69.2	66.4	–	–	–	64.9	68.9	**67.6**
	Aspect	26.8	22.1	22.5	24.7	32.9	21.2	24.3	26.6	25.4	58.8	57.4	**55.9**
	Value	23.9	18.8	19.4	16.7	18.3	17.0	22.7	24.2	23.4	50.9	52.3	**48.9**
	Agent	50.4	51.3	48.9	63.8	67.8	63.2	–	–	–	67.9	72.3	**68.7**
	Patient	34.3	32.2	31.2	23.6	38.3	31.7	–	–	–	32.3	42.9	**35.0**
	Action	23.4	18.8	19.7	21.1	26.0	23.3	31.2	34.1	**32.5**	28.1	39.9	31.4
	Manner	20.3	15.0	15.8	29.4	30.4	28.5	32.7	33.3	33.0	36.2	40.0	**34.7**

Table 8: Tuple prediction results.

To train the sequence taggers and classifiers described above (i.e., the BERT-based static/dynamic classifier, the BERT-BiLSTM-CRF model, and the BERT classifier for predicting an element's value), we use negative cross-entropy as the loss function. The initial learning rate is set to the default value of 0.001. ReLU is chosen as the activation function in the fully-connected layer. Two parameters are tuned using grid search on the development set: dropout (searched out of $\{0.1, 0.2, 0.3, 0.4\}$) and the optimizer (SGD or RMSprop). To train the seq2seq model, we use cross-entropy as the loss function and Adam as the optimizer with an initial learning rate of 0.001. Other parameters are set to their default values.

For all of the models described above, the maximum length of the input is set to 20 (measured in words for English and characters for Chinese). The batch size is set to 8. The number of epochs is searched out of $\{20, 25, 30, 35\}$ using grid search on the development set.

5.3 Results and Discussion

For the first task (static vs. dynamic prediction), our classifier achieves accuracies of 84.9 (English) and 82.1 (Chinese).

Results of the second task (tuple prediction) are shown in Table 8. These results are expressed in terms of recall (R), precision (P), and F-score (F). A few points deserve mention about these results. First, the third approach (Joint CRF-Classifier), where different elements are jointly learned in one model, achieves the best results on the majority of the seven elements for both languages. This indicates that different elements should ideally be identified with different methods (i.e., some using classification and others using extraction). In addition, the fact that it outperforms the other two approaches, where different elements are extracted/generated/predicted by independently-trained models, suggests that there are interactions among the prediction of different elements that can be profitably exploited by a joint model. Second, perhaps not surprisingly, the elements that are most likely to have appeared in the original hyperbolic sentence, including Target, Agent and Patient, have higher F-scores than those elements that are less likely to have appeared in the original sentence, such as Aspect and Value. Third, comparing the first two approaches, the results are rather mixed. Nevertheless, the seq2seq model fails to achieve the best results for any element, which suggests that a generation-based approach may not be suitable for the tuple prediction task.

We conclude this section by reiterating that these are intended to be preliminary results, which we hope can serve as useful baseline results for future work on this task.

6 Conclusion

We introduced the task of deexaggeration and laid the groundwork for further study of this problem by (1) defining a structured representation to represent what is being exaggerated in a hyperbole in a non-hyperbolic manner, (2) annotating the hyperboles in HYPO and HYPO-cn using this representation, (3) conducting an empirical analysis of our annotated corpora, and (4) presenting preliminary results on the task. To stimulate further work on deexaggeration, we make our annotations publicly available.[7]

Acknowledgments

We thank the four anonymous reviewers for their comments on an earlier draft of this paper. This work was supported in part by the National Natural Science Foundation of China (No. 61802167) and the US National Science Foundation (Grant IIS-1528037). Chuanyi Li is the corresponding author.

References

[Chen *et al.*, 2019] I-Hsuan Chen, Yunfei Long, Qin Lu, and Chu-Ren Huang. Metaphor detection: Leveraging culturally grounded eventive information. *IEEE Access*, 7(109):87–98, 2019.

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

[7] The annotations can be downloaded from http://lichuanyi.info/files/papers/Deexaggeration.xlsx.

[Diao *et al.*, 2019] Yufeng Diao, Hongfei Lin, Liang Yang, Xiaochao Fan, Di Wua, and Kan Xua. CRGA: Homographic pun detection with a contextualized-representation: Gated attention network. *Knowledge-Based Systems*, 195:105056, 2019.

[Dong *et al.*, 2019] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*, pages 13042–13054, 2019.

[Felt and Riloff, 2020] Christian Felt and Ellen Riloff. Recognizing euphemisms and dysphemisms using sentiment analysis. In *Fig-Lang@ACL*, pages 136–145, 2020.

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

[He *et al.*, 2019] He He, Nanyun Peng, and Percy Liang. Pun generation with surprise. In *NAACL-HLT*, pages 1734–1744, 2019.

[Hong and Ong, 2009] Bryan Anthony Hong and Ethel Ong. Automatically extracting word relationships as templates for pun generation. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 24–31, 2009.

[Kong *et al.*, 2020] Li Kong, Chuanyi Li, Jidong Ge, Bin Luo, and Vincent Ng. Identifying exaggerated language. In *EMNLP*, pages 7024–7034, 2020.

[Kreuz *et al.*, 1996] Roger J. Kreuz, Richard M. Roberts, Brenda K. Johnson, and Eugenie L. Bertus. Figurative language occurrence and co-occurrence in contemporary literature. In *Empirical Approaches to Literature and Aesthetics*, pages 83–97. 1996.

[Li, 2013] Zhun Li. *A Cognitive Approach to Production Mechanism of Hyperboles: Taking Li Bai's Poem for Exemplification*. Chengdu: Sichuan International Studies University, 2013.

[Liao and Ge, 2014] Yingqiong Liao and Lingling Ge. Hyperbole in humor and its translation. *Journal of Hunan University of Technology Social Science Edition*, 19(4):137–141, 2014.

[Liu and Hwa, 2018] Changsheng Liu and Rebecca Hwa. Heuristically informed unsupervised idiom usage recognition. In *EMNLP*, pages 1723–1731, 2018.

[Liu *et al.*, 2019] Yuanchao Liu, Bo Pang, and Bingquan Liu. Neural-based Chinese idiom recommendation for enhancing elegance in essay writing. In *ACL*, pages 5522–5526, 2019.

[Liu, 2015] Bing Liu. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2015.

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

[Mora, 2009] Laura C. Mora. All or nothing: A semantic analysis of hyperbole. *Revista de Lingüística y Lenguas Aplicadas*, 4(1):25–35, 2009.

[Rivera *et al.*, 2020] Andrés T. Rivera, Antoni Oliver, and Marta Coll-Florit. Metaphoricity detection in adjective-noun pairs. *Procesamiento del Lenguaje Natural*, 64:53–60, 2020.

[Stanivukovic, 2007] Goran V Stanivukovic. "Mounting above the truthe": On hyperbole in English renaissance literature. In *Forum for Modern Language Studies*, volume 43, pages 9–33. Oxford University Press, 2007.

[Stowe *et al.*, 2021] Kevin Stowe, Nils Beck, and Iryna Gurevych. Exploring metaphoric paraphrase generation. In *CoNLL*, pages 323–336, 2021.

[Su *et al.*, 2020] Chang Su, Ying Peng, Shuman Huang, and Yijiang Chen. A metaphor comprehension method based on culture-related hierarchical semantic model. *Neural Processing Letters*, 51(3):2807–2826, 2020.

[Sundararajan and Palanisamy, 2020] Karthik T. Sundararajan and Anandhakumar Palanisamy. Multi-rule based ensemble feature selection model for sarcasm type detection in Twitter. *Computational Intelligence and Neuroscience*, 2020:2860479:1–2860479:17, 2020.

[Troiano *et al.*, 2018] Enrica Troiano, Carlo Strapparava, and Gozde Ozbal. A computational exploration of exaggeration. In *EMNLP*, pages 3296–3304, 2018.

[Tsvetkov *et al.*, 2014] Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. Metaphor detection with cross-lingual model transfer. In *ACL*, pages 248–258, 2014.

[Valitutti *et al.*, 2009] Alessandro Valitutti, Oliviero Stock, and Carlo Strapparava. GraphLaugh: A tool for the interactive generation of humorous puns. In *ACII*, pages 1–2, 2009.

[Yu and Wan, 2019] Zhiwei Yu and Xiaojun Wan. How to avoid sentences spelling boring? Towards a neural approach to unsupervised metaphor generation. In *NAACL-HLT*, pages 861–871, 2019.

[Yu *et al.*, 2018] Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. A neural approach to pun generation. In *ACL*, pages 1650–1660, 2018.

[Zhang *et al.*, 2019] Dongyu Zhang, Hongfei Lin, Xikai Liu, Heting Zhang, and Shaowu Zhang. Combining the attention network and semantic representation for Chinese verb metaphor identification. *IEEE Access*, 7(137):103–110, 2019.

[Zhang, 2016] Haifen Zhang. The exaggerating rhetoric in Mo Yan's novel Sandalwood Penalty. *Journal of Qiqihar Junior Teacher's College*, 1:28–30, 2016.

[Zhao and Lu, 2013] Hong Zhao and Luqianjin Lu. On the actuality and degree of hyperbole. *Journal of Guizhou Minzu University*, 5:90–94, 2013.