Identifying Exaggerated Language

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

Abstract

While exaggeration is one of the most prevalent rhetorical devices, it is arguably one of the least studied in the figurative language processing community. We contribute to the computational study of exaggeration by (1) creating the first Chinese corpus focusing on sentence-level hyperbole detection, with the goal of facilitating a cross-lingual study on this phenomenon, (2) performing a statistical and manual analysis of our corpus, with the goal of gaining insights into the strategies humans employ when creating hyperboles, and (3) addressing the automatic hyperbole detection task with deep learning techniques.

1 Introduction

Recent years have seen a surge of interest in the automatic processing of figurative language in the NLP community, as evidenced by the successful organization of the NAACL 2018 Workshop on Figurative Language Processing. Much of the work on figurative language processing conducted so far, however, 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 vast amount of work on metaphor detection in the past few years was stimulated in large part by the availability of standard evaluation corpora. Progress on the computational study of exaggeration, on the other hand, is hindered by the lack of annotated resources. To our knowledge, HYPO, the first dataset that focuses on exaggeration, was only released in late 2018 (Troiano et al., 2018). HYPO consists of 709 hyperbolic sentences, each of which has a non-hyperbolic version created by manually paraphrasing its hyperbolic counterpart. Given the dataset, Troiano et al. introduced the task of automatic hyperbole detection, where the goal is to determine whether a sentence is a hyperbole.

Given the status quo, our goal is to further the computational study of exaggeration. Specifically, our contributions in this work are three-fold. First, we create HYPO-cn, the first Chinese dataset on exaggeration. HYPO-cn consists of 4762 sentences, of which 2680 are hyperbolic and 2082 are non-hyperbolic. To stimulate research on the computational study of exaggeration, we make HYPO-cn publicly available. We believe that this dataset can complement Troiano et al.'s English dataset and facilitate a cross-lingual study of exaggeration.

Our second contribution involves conducting an empirical analysis of HYPO-cn. We perform two kinds of analysis. First, we conduct a statistical analysis in an attempt to answer various questions involving exaggeration, such as: (1) are there strong lexical indicators of hyperbole; (2) how lexically diverse are the non-hyperbolic versions of a given hyperbolic sentence; and (3) how lexically diverse are the hyperbolic versions of a given non-hyperbolic sentence? Second, we conduct a manual analysis to identify the major strategies used by humans to overstate. We believe our analysis can advance the computational study of exaggeration and allow us to shed light on a number of interesting questions involving exaggeration.

Finally, we perform preliminary experiments on the automatic hyperbole detection task using HYPO-cn. Unlike Troiano et al., who employed only traditional (i.e., non-neural) learners for hyperbole detection, we examine the use of deep learning for model training, with the goal of understanding

whether state-of-the-art learning techniques can offer better results. We show that the best deep learner outperforms the best traditional learner by 11.0% points in accuracy. These results provide suggestive evidence that hyperbole detection is indeed a task that requires a deep understanding of text semantics, as this is what primarily distinguishes a deep learner from a traditional learner.

2 Related Work

2.1 Figurative Language Processing

We begin with an overview of recent work on figurative language processing. For metaphor processing, 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, Hazarika et al. (2018) extract contextual information together with user embeddings in online social media discussions. For idiom processing, Liu and Hwa (2018) identify the intended usage of an idiom in an unsupervised manner, treating possible usages as a latent variable in probabilistic models and training them in a linguistically motivated feature space. Homographic pun detection is addressed by Diao et al. (2019) using a contextualized representation with a gated attention.

The recognition of metaphors and idioms is related to hyperbole detection. Humans sometimes use metaphors and idioms to create hyperbolic sentences (Carston and Wearing, 2011; Zhou and Jiang, 2014). For example, the idiom "Time is money" is a metaphor in which "time" is the noumenon and "money" is the metaphoric object. It is also a hyperbole that overstates the value of time. However, there are differences between metaphor/idiom recognition and hyperbole detection, as many metaphors and idioms are not hyperbolic, such as "The rainbow looks like a bridge".

2.2 Studies on Hyperbole

Compared with other rhetorical devices, hyperbole is less studied. The vast majority of the studies on hyperbole to date have been linguistic rather than computational in nature. Cano Mora (2009), for instance, constructs a taxonomy in which English hyperboles are categorized along two dimensions, quantitative (which involves inflating a quantitative/objective property such as time) and qualitative (which involves inflating a qualitative/subjective property such as emotion). These two dimensions are subcategorized into six semantic fields and 22 subfields. Ferré (2014) shows that at the textual level, a hyperbole can be present in a word or in the interpretation of a certain context.

There are also linguistic studies on exaggeration in Chinese. For instance, Liao and Ge (2014) explore how hyperboles are expressed in the novel "Er Ma". They conclude that hyperboles can be expressed via (1) an upsurge on a semantic scale, which can be qualitative or quantitative, corroborating Cano Mora's findings, 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 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 (i.e., hyperboles not included in the first seven categories). As we will see, some of these strategies are also used to generate sentences in HYPO-cn.

On the computational side, Troiano et al. (2018) create the first annotated English dataset in which every hyperbole has a non-hyperbolic counterpart. They propose the automatic hyperbole detection task, in which they train classifiers to distinguish hyperbolic sentences from non-hyperbolic sentences using traditional machine learners in conjunction with various types of features.

3 Dataset Creation

In this section, we describe the steps involved in the creation of our Chinese dataset, HYPO-cn.

Step 1: Hyperbole Collection

We begin by collecting hyperbolic sentences from two sources: webpages in professional educational websites[1] and linguistics research papers on hy-

[1] www.unjs.com/h/b/148469.html, www.docin.com/p-2191914159.html, www.wnzmb.com/k/kuazhangjudaquan/

perbole in Chinese (Huang, 2010; Zhao and Lu, 2013; Liao and Ge, 2014; Zhou and Jiang, 2014; Zhang, 2016). Specifically, we select 700 sentences from these two sources that have been discussed and determined to be hyperbolic by experts on exaggeration. As a sanity check, we manually go through each of these sentences and verify that all of them are indeed hyperbolic according to the three language-independent criteria of exaggeration summarized by Troiano et al. (2018), namely, the non-literal meaning, the upsurge on a semantic scale, and a connotative trait. We henceforth refer to this set of 700 hyperbolic sentences as S_{hyp}.

Step 2: Non-hyperbole Generation

Next, we hire three native speakers of Chinese to manually produce non-hyperbolic versions of each sentence in S_{hyp}. These annotators are graduate students in NLP (none of them are the authors) and have received a one-hour tutorial on exaggeration from us in which we presented the language-independent criteria of exaggeration described above as well as examples of hyperbolic and non-hyperbolic sentences. After that, each annotator is asked to independently produce a non-hyperbolic version of each (hyperbolic) sentence in S_{hyp} without changing its meaning. By using three annotators, we can examine the extent to which the non-hyperbolic sentences created from the same hyperbole exhibit lexical diversity, and also reduce the possibility that the sentences are biased towards a particular person's style. We henceforth refer to the resulting set of non-hyperboles as S_{common}.

Step 3: Quality Assessment

In order to ensure the quality of the annotations obtained in the previous step, we hire another two annotators to judge if each sentence in S_{common} is a non-hyperbole after giving them the same one-hour tutorial on exaggeration as the other annotators. We delete a sentence if at least one of them thinks that it is hyperbolic or that it does not truly reflect the meaning of its hyperbolic counterpart. After this verification step, 13 non-hyperboles in S_{common} are deleted. For example, the sentence 你嘴里没有实话 (There is no truth in your mouth) is deleted. Although this sentence is often mentioned in daily life, it is overstated as nobody lies all the time. In addition, if two sentences are identical, we delete one of them. Because of this, two sentences are deleted. After this step, every hyperbole in S_{hyp} still has at least one non-hyperbolic counterpart.

Step 4: Hyperbole Generation

Next, we seek to generate more hyperbolic sentences from the non-hyperboles in S_{common}. To avoid the situation where an annotator is being influenced by the original hyperboles (i.e., the sentences in S_{hyp}), we employ another three human annotators who have not seen the sentences in S_{hyp} to manually generate hyperboles after training them in the same one-hour tutorial mentioned above. The sentences in S_{common} that are presented to them are selected as follows: for each sentence s in S_{hyp}, we choose the non-hyperbole version of s from S_{common} that we determine is lexically and syntactically most similar to s. The annotators are required to overstate each non-hyperbole from their own point of view without changing its meaning.

Step 5: Reliability Assessment

Finally, we ask the two annotators involved in Step 3 to judge whether each hyperbolic sentence obtained in the previous step is indeed hyperbolic. Specifically, if at least one of them thinks a sentence is not hyperbolic or does not truly reflect the meaning of its non-hyperbolic counterpart, we will delete it. After this check, 117 sentences are disqualified. For instance, the sentence 一朵朵鲜花红得像血 (Flowers are as red as blood), which is metaphoric, is deleted since many flowers are actually scarlet. As in Step 3, if two sentences are identical, we will only keep one of them. Six sentences are removed because of this.

Overall, more sentences are being deleted in this step than in Step 3. This suggests that writing a qualified non-hyperbole is easier than writing a hyperbole. Although there seems to be more ways to express hyperbole than non-hyperbole, humans often use common expressions (e.g., the same idiom) to overstate, resulting in more repetitive hyperboles than non-hyperboles in our annotation process.

At the end of this process, 700 sentence *sets* are produced, where a sentence belongs to the same set as another sentence if one is a hyperbolic/non-hyperbolic version of the other. In total there are 4762 annotated sentences in HYPO-cn, of which 2680 are hyperbolic and 2082 are non-hyperbolic.

Table 1 shows a sample set of sentences taken from HYPO-cn, where the sentences labeled as 1 are hyperbolic and those labeled as 0 are not. Recall that two sentences in HYPO-cn are in the same set if and only if one is a hyperbolic or non-hyperbolic version of the other. Despite having the same meaning, the sentences within a set exhibit

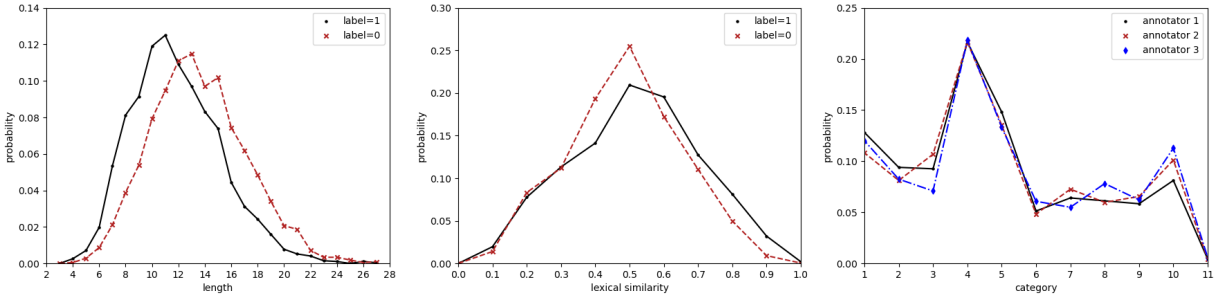

Figure 1: Experiments based on a statistical analysis of HYPO-cn.

Label	Sentence
0	她的腿很长。 Her legs are very long.
	她可以一下子迈上两层台阶。 She can step up two steps once.
	[1] 她的腿一步是我的两步。 Her one step equals my two steps.
1	[2] 她的腿很长，仿佛有两米。 Her legs may be two meters long.
	[3] 她迈开腿，一步就能跨上二楼。 She makes a step and can reach the second floor.
	她的腿比梯子还长。 Her legs are longer than the ladders.

Table 1: An example sentence set in HYPO-cn.

lexical diversity. In fact, even the sentences within the same class can be lexically very different. For instance, there is minimal lexical overlap between the first and second sentences in each class.

4 Corpus Analysis

4.1 Statistical Analysis

We conduct a statistical analysis of HYPO-cn in an attempt to answer several interesting questions about exaggeration.

First, given that hyperbolic sentences may be more descriptive (e.g., compare the first sentence in each class in Table 1), are hyperbolic sentences longer than non-hyperbolic sentences on average? To answer this question, we show in Figure 1(a) the probability distribution of the sentences in HYPO-cn over sentence lengths (as measured by the number of characters). Contrary to our expectation, non-hyperbolic sentences are 2.4 characters longer than hyperbolic sentences on average.

Second, given a random pair of semantically equivalent hyperbolic sentences and a random pair of semantically equivalent non-hyperbolic sentences, which pair is likely to be lexically more diverse? Intuitively there are more ways to express exaggeration, so one would expect the hyperbolic pair to be lexically more diverse than the non-hyperbolic pair. To answer this question, we first compute the cosine similarity of each pair of hyperbolic sentences in the same set as well as the cosine similarity of each pair of non-hyperbolic sentences in the same set, where cosine similarity is computed based on their one-hot word vectors. In other words, the more word overlaps there are between two sentences, the higher their similarity is. We then plot the probability distribution of these sentence pairs over cosine similarity, where cosine similarity is discretized into 10 equal-sized intervals. As we can see in Figure 1(b), both semantically equivalent hyperbolic sentence pairs and semantically equivalent non-hyperbolic sentence pairs are lexically quite diverse: for instance, approximately 20% of the sentence pairs in both categories have a cosine similarity of 0.3 or below, On average, hyperbolic sentence pairs (avg. cosine similarity = 0.46) are lexically less diverse than their non-hyperbolic counterparts (avg. cosine similarity = 0.43). While these results are somewhat contrary to our expectation, the example set in Table 1 may provide hints on why this happened. Specifically, while both the hyperbolic sentences and the non-hyperbolic sentences in Table 1 are lexically quite diverse, a closer inspection should reveal that the average cosine similarity computed over the hyperbolic sentences is higher than that over the non-hyperbolic ones: the topic word 腿 (leg) appears in every hyperbolic sentence but is missing in one non-hyperbolic sentence, and the word 长 (long) appears in two hyperbolic sentences but only one non-hyperbolic sentence. While "two steps" appears in two non-hyperbolic sentences, the two occurrences correspond to different Chinese words and therefore are not considered an overlap.

Third, are there strong lexical indicators of hyperbole? To answer this question, we rank the

Word	WLLR	Word	WLLR
死 (die)	.0486	每个 (each)	.0192
天 (sky)	.0425	瞎 (blind)	.0192
地球 (earth)	.0395	像 (be like)	.0167
般的 (-like)	.0220	一分钱 (a penny)	.0165
命 (life)	.0220	神仙 (immortal)	.0165

Table 2: Ten highest-ranked words computed over the hyperbolic sentences according to WLLR.

words in the hyperbole class by their weighted log-likelihood ratio (WLLR):

$$P(w_t \mid c_j) \log \frac{P(w_t \mid c_j)}{P(w_t \mid \neg c_j)},$$

where w_t and c_j denote the tth word in the vocabulary and the jth class, respectively. Informally, a word w will have a high rank with respect to a class c if it appears frequently in c and infrequently in $\neg c$ (the other class). This correlates reasonably well with what we think an informative word should be.

Table 2 shows the 10 words for the hyperbole class with the highest WLLRs. Looking at each of these words without its context, one may not be able to immediately conclude that the corresponding sentence is hyperbolic. However, one should be able to easily come up with contexts in which these words appear in hyperboles, as some of them are concerned with life and death (死,命) as well as nature and the universe (天,地球), while others indicate the presence of metaphors (般的,像).

4.2 Manual Analysis

The highest-ranked words shown in Table 2 lead us to another question: are there words, phrases, concepts, or even linguistic devices that humans tend to think of and possibly use when they exaggerate?

To answer this question, we ask two native speakers of Chinese who are not involved in any of the previous annotation experiments to perform a manual analysis of the hyperbolic sentences in HYPO-cn. After being trained in the aforementioned one-hour tutorial on exaggeration, they are asked to go over the hyperbolic sentences in HYPO-cn and come up with a way to categorize them, where the categories should shed some light on the strategies humans commonly employ to produce hyperboles. Here, a strategy is broadly construed to include, for example, the use of certain categories of words or phrases, concepts, or linguistic devices. Note that the annotators are *not* asked to go over the non-hyperbolic sentences, as our goal is to identify strategies that are commonly used in a hyperbolic context, rather than those that are used predominantly or even exclusively in hyperbolic sentences.

The annotators come up with 11 categories. Through discussion, they agree on the placement of each sentence into at least one of these 11 categories, which are described below:

1) Quantity concepts. They include (a) expressions with a number or a numeral-measure word combination, such as 两米 (two meters) in sentence [2] and 成千上万 (thousands of), as well as (b) expressions without numbers, such as 无数 (numerous) and 眨眼间 (in the twinkling of an eye). As noted before, the presence of these words/phrases alone is not a sufficient indication of hyperbole: the corresponding sentence is overstated when these expressions are used to quantify an object in a disproportionate, unusual fashion. Nevertheless, quantity concepts are commonly used in hyperboles.

2) Extreme cases. They include (a) completeness and non-exceptionality, such as 全部 (all) and 每 (every), (b) non-existence, such as 一点也不 (not at all), (c) uniqueness, such as 至高无上 (paramount) and 最 (most), as well as (d) boundlessness, such as 无边无际 (boundless) and 无尽 (endless). For example, the use of 所有 (everything) in sentence [4] makes it a hyperbole:

[4] 他是个天才，知道所有的事。
(He is a genius. He knows everything.)

3) Common sayings, including idioms and poems. For example, when describing a stingy person, two annotators use the folk adage 铁公鸡 (iron cock). Unlike the words/phrases in other categories, the idioms used in hyperboles must itself be hyperbolic regardless of the context in which they appear. For example, the idiom 多一事不如少一事 (The less trouble, the better) is not hyperbolic and cannot be employed for exaggeration.

4) Rhetorics. The rhetorical devices that are commonly used in hyperboles include metaphor (sentence [5]), personification (sentence [6]) and synesthesia (sentence [7]):

[5] 那位老先生简直料事如神。
(The old gentleman foretells like a prophet.)

[6] 天气热得连树上的叶子也在喘气。
(It was so hot that the leaves had to gasp for breath.)

[7] 树叶绿得要滴下来了。
(The green color of the leaves is dripping.)

5) Comparison. Hyperbolic sentences that involve a comparison use a reference to highlight

Id	Category	Number	%
1	Quantity concepts	380	14.2
2	Extreme cases	243	9.1
3	Common sayings	268	10.0
4	Rhetorics	686	25.6
5	Comparison	449	16.8
6	Supernatural concepts	160	6.0
7	Desc. about life	176	6.8
8	Desc. of the state of body	234	8.7
9	Desc. about nature	201	7.5
10	Fictitious scene	298	11.1
11	Impossible ordering	17	0.6

Table 3: Eleven categories of strategies employed by humans when creating hyperboles.

the characteristic of an object. Sentence [8] is an example that makes a comparison between "mind" and "sky".

[8] 他的心胸比天空宽阔。
(His mind is wider than the sky.)

Sky is commonly known to be boundless, so the contrast in the sentence underlines his generosity.

6) Description about supernatural concepts. Sentence [5] stresses how clever the old gentleman is. As we know, prophets, gods, and immortals are among the most powerful and intelligent beings, so drawing a connection between a human being and a supernatural being is a way to overstate.

7) Description about life. This semantic field includes (a) the concept of bringing/destroying life, such as 生命 (life), 重生 (reborn), and 要命 (fatal), (b) physical health, such as 病 (sick), as well as (c) mental state, such as 发疯 (crazy) and 精神病 (psychosis). Sentence [9] is hyperbolic because while the urge is annoying, it can never kill you.

[9] 他们的催促要索命。
(Their urge is killing.)

8) Description about the state of the human body. This category involves sentences that express exaggeration via describing an unusual state of the human body or organ. For example, in sentence [10], when overstating the word 看着 (stare at), the depiction of the person's eyes is used to highlight the great deal of concentration.

[10] 他看着那位小姐，大眼珠险得突破眼眶。
(He stares at the young lady. His big eyes are breaking through the orbit.)

9) Description about nature. This semantic field includes entities in nature and natural phenomena with distinctive features, such 闪电 (lightning), 地球 (earth), and 南极 (Antarctica). For example, sentence [9] uses 天空 (sky) to describe "wide" as the sky is known to be extremely vast.

10) Fictitious scene. Sometimes a human employs an imaginary scene to overstate his/her point. Sentence [3], for instance, describes a scene where "she" reaches the second floor in one step in order to highlight how long her legs are.

11) Impossible ordering. This category of sentences describes a situation in which the sequence of events involved did not take place in a possible order, as in sentence [11]:

[11]在娘肚子里我就会抽烟了。
(I learned to smoke before I was born.)

Table 3 shows the number and percentage of sentences in HYPO-cn that involve each strategy. As mentioned earlier, these categories are not disjoint. As we can see, Rhetorics is the largest category, whereas Impossible Ordering is the smallest.

An interesting question is: do humans employ the same strategy for exaggeration when trying to rewrite a non-hyperbolic sentence? Recall from Section 3 that in Step 4 of our corpus creation procedure, three hyperbolic sentences are created by having three annotators rewrite a non-hyperbolic sentence. To answer the above question, we compute statistics on how often our three annotators employ the same strategy when writing hyperboles using these non-hyperbolic sentences.[2] At the outset, we do not expect high agreement, as more than one strategy can be assigned to a given sentence.

There are 53 sets (8.9%) in which all three annotators employ the exact same strategy. Excluding these 53 sets, we have 41 sets (6.9%) in which all three annotators have at least one strategy in common (e.g., two of them employ strategy 1 while the third employs strategies 1&2). Among the remaining sets, there are 233 (39.0%) in which two annotators employ the exact same strategy and another 93 (15.6%) in which two annotators have at least one strategy in common. Finally, there are 178 sets (29.8%) in which the three annotators all employ different strategies.

It is somewhat surprising that in as many as 30% of the sets the annotators use different strategies. To gain insights into the reason, we show in Table 4 a set that belongs to this category, where the original (non-hyperbolic) sentence is in the first row. As we can see, the three annotators employ three different strategies: 5 (comparison), 1 (quantity concepts), and 2 (extreme cases).

[2]Recall that some hyperbolic sentences are discarded in Step 5. We compile our statistics based on only the 598 sets where none of the three hyperbolic sentences are discarded.

Sentence	Id
凭你的本事，我没有办法可以瞒住你。 Given your ability, I have no way to fool you.	–
凭你的本事，瞒住你比造火箭还难。 Given your ability, fooling you is more difficult than making the rocket.	5
凭你的本事，我使出十八般武艺都瞒不住你。 Given your ability, I cannot fool you even if I employ eighteen skills.	1
凭你的本事，瞒住谁也不可能瞒住你。 Given your ability, no one can fool you.	2

Table 4: Example set in which all three annotators employed different exaggeration strategies.

Recall that the reason for our employing multiple annotators in Section 3 is to ensure that the resulting corpus reflects diverse ways of expressing hyperbole. A relevant question is: how different are the annotators in terms of the strategies they choose to express hyperboles? To answer this question, we show in Figure 1(c) the probability distribution of the sentences produced by each annotator over the 11 strategies. While the annotators differ in terms of how often they employ a particular strategy, the three plots exhibit similar patterns. These results seem to suggest that the corpus would not have been severely biased in terms of the way the hyperboles were expressed even if it had been annotated by just one person.

5 Automatic Hyperbole Detection

Next, we present preliminary results on the automatic hyperbole detection task. We cast it as a supervised binary classification problem where we predict whether a sentence is hyperbolic or not.

5.1 Traditional Learning Algorithms

As baseline systems, we employ those used by Troiano et al. (2018), who adopt a set of traditional machine learning algorithms encapsulated in the Sklearn library (Pedregosa et al., 2011) using the default learning setting, including logistic regression (LR), k-nearest neighbor algorithms (KNN), Naïve Bayes (NB), decision tree learners (DT), support vector machines (SVM), and Latent Dirichlet Allocation (LDA), to train classifiers for determining if a sentence is a hyperbole or not.

We employ the aforementioned learners for model training in conjunction with two types of features, hand-crafted features and embedding features, as described below.

The *hand-crafted features* are taken from those described in Troiano et al. (2018). More specifically, we reimplement four of the five hand-crafted features used by Troiano et al., namely Unexpectedness (how coherent a word is with the rest of the discourse), Polarity (the sentiment of the sentence), Subjectivity (whether the sentence is objective or not), and Emotional Intensity (the sentiment strength of the sentence). To be specific, we compute Unexpectedness with the pretrained Skip-gram vectors provided by Google (Mikolov et al., 2013) and the Directional Skip-gram embeddings provided by Song et al. (2018), and Polarity with the SnowNLP library[3] and HowNet[4]. However, we are unable to implement the Imageability feature, which encodes the degree to which a word evokes a mental image. The reason is that Troiano et al. computed this feature based on the imageability ratings of the MRC psycholinguistic database, but such a resource is absent for Chinese. According to Troiano et al., the quantitative criteria for hyperboles are encoded partially by Unexpectedness, whereas the qualitative criteria are encoded by Polarity, Subjectivity, and Emotional Intensity. We will henceforth refer to this feature set as TF.

Embedding features are features derived from word embeddings. These features have recently been used extensively in various NLP tasks. We experiment with three types of pre-trained word embeddings: (1) the 300-dimensional Skip-gram representations, (2) the 200-dimensional Directional Skip-gram embeddings, and (3) the 768-dimensional contextualized embeddings trained based on BERT (Devlin et al., 2019), which we obtain by feeding the input sentence into Cui et al.'s (2019) implementation of BERT for Chinese. We generate the embedding features for a given sentence by averaging the embeddings of its constituent words. We will refer to the resulting feature sets produced via Skip-gram, Directional Skip-gram, and BERT as SG, DS, and BE, respectively.

5.2 Neural Network Setting

Troiano et al. (2018) employ only traditional learners in their experiments. A natural question is: will deep learners offer better performance on the hyperbole detection task? To answer this question, we employ the two commonly used deep learners in NLP, namely CNN and LSTM, as realized in the Keras API (Chollet et al., 2015).

For CNN, we use one convolutional layer and

[3] https://github.com/isnowfy/snownlp

[4] www.keenage.com/html/c_bulletin_2007.htm

	Model	TF	SG	SG+TF	DS	DS+TF	BE	BE+TF	Words	Words+TF
Traditional learners	LR	58.9	66.5	67.2	74.4	73.7	73.9	74.3	–	–
	KNN	57.2	58.6	59.6	60.8	60.4	63.5	63.2	–	–
	NB	59.3	60.6	61.1	62.9	62.5	61.5	61.8	–	–
	DT	54.6	55.8	55.3	59.0	59.2	58.3	58.5	–	–
	SVM	58.2	67.9	68.0	74.2	74.3	74.1	73.9	–	–
	LDA	59.0	66.0	67.0	72.4	72.6	70.9	71.1	–	–
Deep learners	CNN	–	80.7	81.8	83.6	84.1	82.1	81.6	–	–
	LSTM	–	82.8	82.6	84.7	**85.4**	83.2	83.3	–	–
	BERT	–	–	–	–	–	–	–	78.5	78.3

Table 5: Ten-fold cross-validation accuracies on automatic hyperbole detection.

max-pooling. For both CNN and LSTM, we experiment with the three types of word embeddings described in the previous subsection to represent the words in the input sentence and employ ReLU as the activation function with a mini-batch size of 32. The dropout rate and the number of epochs are tuned to maximize accuracy on held-out development data using grid search.[5] We use negative cross-entropy as the loss function and SGD as the optimizer with an initial learning rate of 0.001.

For comparison purposes, we fine-tune a BERT model pre-trained on Chinese (Cui et al., 2019) for our task as follows. If the input is only composed of the sentence to be classified, then we pass it to the encoder, feed the embedding of [CLS] token (in the last layer) to a task-specific classification layer, and jointly fine-tune the model parameters of BERT and the classifier. If the hand-crafted features described above are additionally used as input, we simply concatenate the corresponding feature vector with the embedding of the [CLS] token and use the resulting vector for fine-tuning. We tune the dropout rate and the number of epochs in the same way as we did for CNN and LSTM.

5.3 Experimental Setup and Results

Since Chinese has no space between words, we use the Jieba library[6] for word segmentation. We remove the stopwords from each sentence and report 10-fold cross-validation results in all experiments. In each fold experiment, we use eight folds for model training, one fold for parameter tuning, and one fold for testing. Each fold contains exactly 70 sentence sets.[7] We report performance in terms of accuracy. Note that the majority baseline, which classifies every test instance as hyperbolic, has an accuracy of 56.2%.

Experimental results are shown in Table 5. In addition to using the two types of features (hand-crafted features and embedding features) in isolation, we also use them in combination. For the traditional learners, we simply concatenate the two sets of features. For the deep learners, the hand-crafted features are concatenated with the output of the encoder in the dense layer.

Several points deserve mention. First, using only the hand-crafted features, we obtain mixed results. DT, the worst-performing learner on this task, underperforms the majority baseline. While NB yields the best performance, it only achieves an accuracy of 59.3.

Second, using only the embedding features always yields better results than using only the hand-crafted features, regardless of which traditional learner and which type of embedding are used. Nevertheless, LR, SVM and LDA seem to make more effective use of the embedding features than the remaining learners, and among the three types of embeddings, DS generally offers the best results while SG generally offers the worst results.

Third, using both the hand-crafted features and the embedding features is not necessarily better than using only the embedding features. Overall, the results are rather mixed: in the presence of the embedding features, the hand-crafted features only offer slightly improved performance in the majority of the cases.

Fourth, the two deep learners, CNN and LSTM, achieve substantially better results than the traditional learners: the worst deep learning-based system outperforms the best traditional learning-based system by at least 3.9% points, and the best deep learning-based system outperforms the best traditional learning-based system by 11.0% points. The best result, 85.4%, is achieved by using LSTM in

[5] Dropout rate: we tried values from 0.1 to 0.4 in increments of 0.1; number of epochs: we tried 20, 25, and 30.

[6] https://github.com/fxsjy/jieba

[7] All the sentences in the same set will appear in the same fold. This setup could minimize the lexical similarity across different folds, as the sentences in the same set are likely to be lexically more similar than those that appear in different sets.

Features	CNN		LSTM	
	NH	H	NH	H
SG	71.5	80.8	76.3	82.9
SG+TF	74.2	81.4	76.1	82.2
DS	78.1	82.8	79.2	84.2
DS+TF	79.1	83.9	79.6	84.5
BE	75.8	81.6	77.8	82.6
BE+TF	74.3	81.1	78.0	82.5

Table 6: F-scores achieved by different models on the hyperbolic (H) and non-hyperbolic (NH) classes.

conjunction with DS+TF.

Finally, as seen in the last two columns of Table 5, whether or not hand-crafted features are used, fine-tuning BERT yields results that are better than those produced by the traditional learners but not as strong as those produced by the deep learners.

To better understand how well the models identify the hyperbolic sentences and the non-hyperbolic sentences, we show in Table 6 the F-scores achieved on the hyperbolic (H) class and the non-hyperbolic (NH) class by two of the best learners, CNN and LSTM. Two points deserve mention. First, the F-scores on H are better than those on NH. This is perhaps not surprising, since there are more hyperbolic sentences than non-hyperbolic sentences in the corpus. Second, the DS results are better than the SG results and the BE results because of better identification of sentences in both classes, even though the improvements on the non-hyperbolic sentences are generally more pronounced than those on their hyperbolic counterparts.

5.4 Error Analysis

We perform an error analysis on the best-performing model, namely LSTM with DS+TF, and observe the following major error categories:

Failure to understand context. A word like 最 (most) is sometimes used in hyperboles, but not in the context in sentence [12]. Without understanding the context in which indicators like this appear, the model got confused and misclassified the corresponding sentence as hyperbolic.

[12] 她是家里孩子中最聪明的。
(She is the most intelligent child in the family.)

Lack of background knowledge. To properly understand sentence [13], one should have the background knowledge that Four Books and Five Classics are masterworks in China. Without such information, the model failed to understand the sentence and misclassified it as non-hyperbolic.

[13] 老太太的经验就是我们的四书五经。
(The old lady's experience is our Four Books and Five Classics.)

Lack of commonsense knowledge. Sentence [1] is non-hyperbolic because the fact that "her one step equals my two steps" is not anything that would be surprising to anyone. However, if one changes the number from "two" to "100", then the resulting sentence becomes hyperbolic because in reality it is not possible that one person's step would equal another person's 100 steps. Hence, the determination of whether a sentence is hyperbolic or not often requires the commonsense knowledge of whether the word/phrase used to describe an object is out of normal range or not. This kind of knowledge is currently missing in our system.

6 Conclusion

We presented an empirical study of exaggeration, which is one of the most prevalent and yet one of the least studied rhetorical devices from a computational perspective. Our contributions lie in (1) creating a Chinese corpus focused on exaggeration, (2) identifying different strategies used by humans for exaggeration and (3) showing that deep learners substantially outperform traditional learners on automatic hyperbole detection. The statistical and manual analyses of our corpus, which is absent from other computational studies on exaggeration, have shed light on various interesting questions about this rhetorical device. To stimulate research on this topic, we make HYPO-cn publicly available.[8] In future work, we plan to use HYPO and HYPO-cn to conduct a cross-lingual study on whether there are differences in the way exaggeration is expressed in English and Chinese.

Acknowledgments

We thank the three anonymous reviewers for their detailed and insightful comments on an earlier draft of the paper. This work was supported in part by the National Natural Science Foundation of China (No. 61802167) and the US National Science Foundation (Grants IIS-1528037 and CCF-1848608). Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of the funding agencies.

[8] HYPO-cn can be downloaded from http://lichuanyi.info/paper/chinese_hypo.txt.

References

Laura Cano Mora. 2009. All or nothing: A semantic analysis of hyperbole. *Journal of Language and Applied Language*, 4:25–35.

Robyn Carston and Catherine Wearing. 2011. Metaphor, hyperbole and simile: A pragmatic approach. *Language and Cognition*, 3(2):283–312.

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

Francois Chollet et al. 2015. Keras. https://github.com/fchollet/keras.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for Chinese BERT. *arXiv:1906.08101*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 4171–4186.

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

Gaëlle Ferré. 2014. Multimodal hyperbole. *Multimodal Communication*, 3(1):25–50.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. CASCADE: Contextual sarcasm detection in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848.

He He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1734–1744.

Guan Huang. 2010. Application of hyperbole in relationship from pragmatics. *Journal of Chongqing College of Education*, 23(4):113–116.

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

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

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

Changsheng Liu and Rebecca Hwa. 2018. Heuristically informed unsupervised idiom usage recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1723–1731.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of 27th Annual Conference on Neural Information Processing Systems*, pages 3111–3119.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 175–180.

Enrica Troiano, Carlo Strapparava, and Gozde Ozbal. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258.

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

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

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

Jiajia Zhou and Bin Jiang. 2014. Relationship between metaphor and hyperbole. *Journal of Chongqing Three Gorges University*, 30(154):113–116.