ConFiguRe: Exploring Discourse-level Chinese Figures of Speech

Dawei Zhu   Qiusi Zhan   Zhejian Zhou   Yifan Song
Jiebin Zhang   Sujian Li*
School of Computer Science, Peking University
Key Laboratory of Computational Linguistics, MOE, Peking University[†]

Abstract

Figures of speech, such as metaphor and irony, are ubiquitous in literature works and colloquial conversations. This poses great challenge for natural language understanding since figures of speech usually deviate from their ostensible meanings to express deeper semantic implications. Previous research lays emphasis on the literary aspect of figures and seldom provide a comprehensive exploration from a view of computational linguistics. In this paper, we first propose the concept of figurative unit, which is the carrier of a figure. Then we select 12 types of figures commonly used in Chinese, and build a Chinese corpus for Contextualized Figure Recognition (ConFiguRe). Different from previous token-level or sentence-level counterparts, ConFiguRe aims at extracting a figurative unit from discourse-level context, and classifying the figurative unit into the right figure type. On ConFiguRe, three tasks, i.e., figure extraction, figure type classification and figure recognition, are designed and the state-of-the-art techniques are utilized to implement the benchmarks. We conduct thorough experiments and show that all three tasks are challenging for existing models, thus requiring further research. Our dataset and code are publicly available at https://github.com/pku-tangent/ConFiguRe.

1 Introduction

Figures of speech, also known as rhetoric figures or figurative languages, are a ubiquitous part of spoken and written discourse. These rhetorical techniques, such as metaphor, irony and parallelism, greatly enrich the expression of human languages (Roberts and Kreuz, 1994). They intentionally deviate from the literal meaning of language to provide deeper semantic expressiveness, therefore

---

*Corresponding author

[†] Author emails:{dwzhu, zhanqiusi, zhouzhejian, yfsong, zhangjiebin, lisujian}@pku.edu.cn

Figure 1: An annotated case of ConFiguRe. We present English translation in the second block for reading convenience. Clause groups in *italic* are figurative units, with the specific figure type denoted in subscription.

posing a big challenge to natural language understanding.

Linguists have a long history of studying rhetoric figures, extensively analyzing their use in literature, culture and psychology (Zhang, 1963; Wilks, 1975; Drew and Holt, 1998; Group, 2007; Shapin, 2012). These works mainly focus on collecting qualitative evidence. Figures of speech have also drawn attention from the NLP community. Many downstream applications could be improved, should figures be precisely identified and carefully dealt with. For example, a faithful translation should adapt the metaphors used in the source language to an authentic expression in the target language, and sentiment analysis should benefit from the correct identification of irony and sarcasm.

Despite its significance, a comprehensive study of identifying rhetoric figures from discourse remains under-explored. Previous works mainly emphasize specific figures, such as metaphor (Steen, 2010; Fass, 1991; Dodge et al., 2015; Su et al., 2020) and sarcasm (Khodak et al., 2017; Davidov et al., 2010; Wallace et al., 2014; Lee et al., 2020),

or identify rhetoric figures at a token or sentence level (Wen et al., 2019; Chen et al., 2021). However, in real-world settings, figurative languages are hidden in a long context and the potential figure type is usually unknown. Motivated by this, we construct a comprehensive dataset of 12 commonly used figures in Chinese, and include discourse fragment as context for each instance.

First we describe our guideline towards constructing a rhetoric corpus, which is devised upon linguistic theory and existing reading practice. The design of the guideline is oriented with two key questions: (1) What is the language carrier of a specific figure? (2) Which figures should be included in our corpus? For the first question, we propose the concept of *figurative unit*——the smallest continuous clause sequence containing a complete expression of a specific figure. For the second question, we approach it from a linguistic view. Linguistically, figures can be divided into two main groups: *schemes* and *tropes*. *Schemes* reflect a deviation from the ordinary pattern or arrangement of words, while *tropes* involve deviation from the ordinary and principal signification of words (Corbett, 1999). Following previous work of Chinese linguists (Zhang, 1963) and existing reading-comprehension practice, we select 7 tropes and 5 schemes commonly used in Chinese in our work.

Following the aforementioned guideline, we leverage human annotation to build ConFiguRe, a Chinese corpus for Contextualized Figure Recognition. Each instance in ConFiguRe includes a discourse fragment with several annotated figurative units attached to it. An annotated instance of ConFiguRe is illustrated in Figure 1. In this instance, the fragment includes 12 Chinese clauses and 2 figurative units. Clause 4 alone is a figurative unit labeled with the *metaphor* type, while the latter 4 clauses constitute another figurative unit labeled *parallelism*. ConFiguRe is, to the best of our knowledge, the first rhetoric dataset that involves both extracting a figurative unit from the discourse-level context and classifying this unit into the right figure type. In comparison, previous datasets mainly focus on detecting a specific figure from a given sentence (Joshi et al., 2016; Khodak et al., 2017).

For benchmark settings, we design three tasks based on ConFiguRe, namely *figure extraction*, *figure type classification*, and *figure recognition*. We deploy state-of-the-art models as baselines, and reveal that all three tasks remain challenging through thorough experiment. We also conduct subsidiary experiments to explore future directions for these tasks, which will contribute to the research of this area.

To sum up, our main contribution is threefold:

- We design the guideline towards constructing a discourse-level rhetoric corpus based on linguistic theory, and propose the concept of *figurative unit* as the basic element for analysis.

- We construct ConFiguRe, a human-annotated Chinese corpus for contextualized figure recognition, which includes 12 most frequently used figure types. Upon this, we design three tasks as benchmarks: *figure extraction*, *figure type classification*, and *figure recognition*.

- We implement models based on recent state-of-the-art techniques as baselines, and conduct thorough experiments and analysis. We find that all three tasks on ConFiguRe are challenging with a lot of room to improve.

2 Related Work

Over the last decade, automated detection of figurative languages has become a popular topic, and a considerable number of datasets in this area have been constructed. These datasets can be roughly divided into two categories: the first is to extract the span of a specific figure given a sentence and its context (span extraction); the second is to determine whether a sentence is figurative (sentence classification). For span extraction, it is usually accomplished by marking out tokens carrying the target figure. These works include VUA (Joshi et al., 2016) and the NTU Irony Corpus (Tang and Chen, 2014). For sentence classification, there are two lines of relevant research. The first is a binary sentence classification task for determining whether a given sentence is figurative, e.g. SARC (Khodak et al., 2017) for sarcasm, and the Chinese rhetoric question dataset built by Wen et al. (2019). The second is to classify a figurative sentence into its corresponding figure type (multi-classification), such as the Chinese multi-label rhetoric dataset constructed by Chen et al. (2021) for joint rhetoric and emotion identification.

On the basis of existing datasets, a series of methods have been developed, ranging from feature engineering (Bulat et al., 2017; Köper and Schulte im

Walde, 2017; Tsvetkov et al., 2014) to neural networks (Liu et al., 2018; Chen et al., 2021; Mu et al., 2019; Dankers et al., 2020; Joshi et al., 2017; Leong et al., 2020, 2018). However, perhaps limited by the fact that existing datasets only target at one figure, previous works mainly deal with specific figure types. A comprehensive study of a collection of figure types remain under-explored.

One further observation is that previous datasets are rarely shipped with wider contextual information. They mainly approach the subject from a token-level or sentence-level perspective. That being the situation, a series of works have pointed out that leveraging contextual information is beneficial in figure detection (Dankers et al., 2020; Mu et al., 2019; Jang et al., 2015; Joshi et al., 2017, 2015). Joshi et al. (2015) proposed to use text incongruity from linguistic theory for sarcasm detection. Dankers et al. (2020) used a general and a hierarchical attention mechanism for modeling discourse, improving SOTA for the 2018 VU Amsterdam (VUA) metaphor identification shared task (Leong et al., 2018) by 6.4 F1-scores. Other contextual clues such as author context (Bamman and Smith, 2015; Ghosh and Veale, 2017), multi-modal context (Schifanella et al., 2016; Castro et al., 2019), conversation context (Joshi et al., 2016; Ghosh et al., 2017), have also been proven to improve figure detection.

In this paper, we propose Chinese Corpus for Contextualized Figure Recognition (ConFiguRe), which, to the best of our knowledge, is the first comprehensive corpus that includes more than 10 commonly used figures in Chinese, with relevant discourse fragments serving as contextual information for each instance. Our dataset can be used for both figure extraction and figure type classification.

3 Corpus Construction

Previous rhetorical studies (Group et al., 2007; Harris et al., 2018) have contributed many useful annotation paradigms. However, These annotation paradigms are primarily customized for one specific figure type, or tailored to scheme figures that present themselves with strict patterns, which are quite inapplicable in our setting. In this section, we present our self-devised annotation guideline to build ConFiguRe with 7 trope figures and 5 scheme figures.

1. Metaphor/Simile: 阳光像金子一样珍贵，(Sunshine is as precious as gold)
2. Parallelism: 童年是小草的芽儿，充满了生机；童年是早晨的太阳，充满了活力；童年是清润的雨水，充满了欢乐。(Childhood is the bud of grass, full of vitality; childhood is the morning sun, full of vitality; childhood is the rain, full of joy.)

Figure 2: Two examples for figurative units. First is a *Simile* unit containing one single clause, the second is a *Parallelism* unit composed of six clauses.

3.1 Annotation Guideline

Figurative Unit For the convenience of analyzing figures, we first posit the concept of *figurative unit* as the basic language carrier of a figure. A *figurative unit* is defined as "the smallest continuous clause sequence carrying a complete expression of a specific figure". [1] The intuition is that, under most circumstances, a figure instance only comprises a limited portion of a sentence. Hence, we prefer clauses as elementary constituent of a figure, as it is more fine-grained than a sentence. In Figure 2, we demonstrate two examples of figurative units with their corresponding figure types. In the first example, a single clause is a figurative unit carrying the figure *Simile*; while in the second example, six clauses together form a figurative unit of *Parallelism*.

Figure Types Upon selection of figure types, we refer to the linguistic categories from English (Burton, 2016) and Chinese (Zhang, 1963). We choose the most frequently used ones in written literature. It should be noted that while some figures are widely used in English, they do not have exact counterparts in Chinese, therefore excluded from our dataset. At the same time, we avoid choosing the figures whose identification may involve deep semantic background, e.g. Pun, Paradox, and leave them for future study. Specifically, we adopt the following 12 figures: *Metaphor/Simile[2]*, *Personification*, *Metonymy*, *Hyperbole*, *Irony*, *Synaesthesia*, *Rhetorical question*, *Parallelism*, *Duality*, *Repetition*, *Antithesis*, *Quote* (For detailed description of each figure, see Appendix A.1). For simplicity, we temporarily use the first four or five characters as

---

[1] It is debatable about the definition of a Chinese clause. Here for simplicity, we define a clause as a text span separated by the separator punctuations. A full list of punctuations is included in Appendix A.2.

[2] In Chinese, these two figures are combined as a single figure type which means a comparison made by referring to one thing as another.

the abbreviation of each figure in this paper. Note that we do not aim at including all figure types in our research, but hope to recognize the commonly-used figures out of the discourse.

In linguistics, figures are divided into two main groups: *schemes* and *tropes*. *Schemes* reflect a deviation from the ordinary pattern or arrangement of words, while *tropes* involve deviation from the ordinary and principal signification of words (Corbett, 1999). Among the 12 commonly used figures, the first 7 are *tropes* and the last 5 are *schemes*. We adopt this categorization to facilitate following analysis below.

3.2 Data Collection

Though a common technique in communication, figurative languages are actually sparsely distributed in main-stream corpora, consisting of news, dialogs, etc. For a higher proportion of figurative languages, we build our dataset on a collection of literary works. 98 Chinese literary works, most of which are novels and proses, are collected from publicly available resources.

We then divide each literary work into fragments comprising of whole paragraphs, while also ensuring that each fragment contains no more than 500 words. In this way, we assert that the discourse is neither too short to miss important contextual information nor too long to distract annotators' attention. In total, we obtain 12,976 fragments. These fragments are later presented to human annotators for manual annotation.

3.3 Human Annotation

We recruit 17 annotators whose native language is Chinese. These annotators are all well-educated, mostly majoring in linguistics. We divide them into two groups as *fragment annotators* and *figurative unit annotators*, provide instructions with detailed definitions and examples of each figure type, and train them for figure annotation before setting out on the full dataset. The annotation process is carried out coarse-to-fine in two stages.

The first stage is for coarse classification, where 5 fragment annotators are asked to classify fragments as figurative or not. Each fragment is annotated by one annotator only once, since we suppose classifying a discourse fragment as figurative or not is a binary classification task that is relatively easy.

The second stage involves 12 figurative unit annotators. Each figurative fragment is presented to

Split	# frag.	wd./frag.	cls./frag.	# figUnit.	figUnit./frag.
train	2934	419.7	41.9	6254	2.1
valid	419	417.5	42.1	886	2.1
test	839	419.2	41.3	1870	2.2
Total	4192	419.4	41.8	9010	2.1

Table 1: Statistics of train, valid and test set in our ConFiguRe. *frag.*, *wd.*, *cls.*, *figUnit.* is short for fragment, word, clause, figurative unit, repectively.

two annotators, who are independently required to extract figurative units (one fragment can contain several units) and assign the corresponding figure types. We make sure that each figurative unit can only be labeled with one figure type. Additionally, conflict-solving strategies are devised in case of inconsistent labeling. For example, suppose annotators A_1 and A_2 respectively extract figurative units u_1 and u_2, the inconsistency can be roughly divided into 3 categories: 1) u_1 and u_2 are figurative units that do not overlap, in this case we keep both of them. 2) u_1 is a subset of u_2, and they are assigned the same figure type, in this case we choose u_1 as the gold annotation, since figurative unit is defined to be the smallest clause group containing a complete expression of a specific figure. 3) u_1 overlaps with u_2 but is neither a superset or subset, or they are assigned with different figure types, we regard this case as the most complex one and ask another annotator to make a decision based on the annotation results of annotators A_1 and A_2. Besides, if an annotator extracts several figurative units in one go, inconsistency can be solved similarly.

In addition to identifying the figurative units and their types, we also ask the annotators to corroborate their judgement with evidence (e.g. strong feature words denoting the figure) which may benefit future work.

The main reason of designing this coarse-to-fine annotation process is that the work of fragment annotators can narrow down the context of a figure and make the figurative unit annotators pay more attention to identifying the boundary and type of a figure.

3.4 Dataset Analysis

Through human annotation, we present ConFiguRe with 4,192 figurative fragments and 9,010 figurative units. Note that each instance in ConFiguRe is a discourse fragment carrying figurative languages.

Figures	# figurative unit			words/unit	clauses/unit
	train	valid	test		
Meta.	2226	306	683	18.5	1.55
Pers.	768	94	243	19.5	1.64
Meto.	411	69	123	18.3	1.68
Hyper.	473	72	145	19.5	1.75
Irony	24	4	5	27.5	2.36
Syna.	18	6	10	22.9	1.76
Rheq.	868	115	202	19.5	2.07
Para.	291	41	99	37.1	3.90
Dual.	272	37	76	16.8	2.24
Repe.	382	57	105	16.6	2.98
Anti.	179	23	62	28.7	2.73
Quote	342	63	117	33.3	3.90
Total	6254	886	1870	20.7	2.05

Table 2: Statistics of each figure in train, valid and test set. For simplicity, we temporarily use the first four or five characters as the abbreviation of each figure in this paper. First seven figures are *tropes*, last five figures are *schemes*.

Each fragment may contain one or more annotated figurative units. Table 1 demonstrates basic statistics of the train, valid and test set, according to the split proportion of 7:1:2.

Detailed information of each figure type is rendered in Table 2. We can see that *metaphor/simile*, *personification*, and *rhetoric question* occur more frequently than others, while *synaesthesia* and *irony* are extremely hard to gather in our corpus. For these two figure types, we may collect more instances in the future. From Table 2, we can also see that the average number of clauses for *schemes* is consistently larger than that of *tropes*. This can be attributed to the fact that *schemes* are usually only identifiable in a sequence of clauses, while *tropes* can be directly expressed within one clause. Interestingly, the average word number of *irony* figures is comparable to or even longer than many scheme figures. A priori is that the expression of an *irony* requires more contextual information. Scheme figures *duality* and *repetition* are relatively short, probably due to the fact that *duality* often leverages short clauses and *repetition* evinces at the word level.

4 Task Definition and Baseline Models

Based on ConFiguRe, we propose three benchmark tasks. Task 1 is figure extraction, which extracts figurative units from input text. Task 2 is figure type classification, which classifies a figurative unit into the corresponding figure type. Task 3 is figure recognition composed of previous two tasks, which extracts figurative units from input text and determines their corresponding figure types simultaneously.

Although there have been a lot of methods in rhetoric detection, we find that these methods mostly catered for one specific figure (Leong et al., 2020; Ghosh et al., 2020), and therefore not quite consistent with our task settings. For baseline models, we adopt self-designed approaches backboned with the state-of-the-art RoBERTa model. We give a formal definition for each task and introduce the corresponding baseline models as follows.

4.1 Task 1: Figure Extraction

Given a discourse fragment $\mathcal{D}$ comprising m clauses and n words, let $\mathcal{D} = \{c_i\}_{i=1}^{m}$, where $c_i = \{w_j\}_{j=1}^{n_i}$ refers to the i-th clause consisting of n_i words, and $\sum_{i=1}^{m} n_i = n$. Task 1 aims to extract figurative units $u_1 = (c_{b_1}, \ldots, c_{e_1}), \cdots, u_k = (c_{b_k}, \ldots, c_{e_k})$ from $\mathcal{D}$, where k is the number of figurative units in $\mathcal{D}$, u_i is the i-th figurative unit, and the subscripts b_i and e_i denote the beginning and ending positions, respectively. Note that each figurative unit is the smallest continuous clause sequence carrying a complete expression of a specific figure. (See Section 3.1 for detailed explanation of clause and figurative unit)

To perform figure extraction, we first use RoBERTa encoder and perform clause-wise mean pooling to obtain contextualized representations $\boldsymbol{h_c} = (h_{clause_1}, \ldots, h_{clause_m})$ for each clause, where m is the number of clauses. Based on this, we design the following baselines for Task 1:

FESeq FESeq is implemented by modeling figure extraction as a clause-level sequence labeling task. Following traditional settings, the labels "B" "I" and "O" are assigned to each clause when it is the first clause of a figurative unit, inside but not the first of a figurative unit, and not in a figurative unit accordingly. A classification layer is added on top of RoBERTa, casting hidden representations into 3-dimensional logits for "B", "I" or "O".

FECRF Since Conditional Random Field (Lafferty et al., 2001) is a common strategy in sequence labeling tasks, we design the FECRF model by adding a CRF layer on top of FESeq model.

FESpan FESpan is developed by modeling figure extraction as a clause-level binary span tagging task. It adopts two binary MLP classifiers upon

encoder to detect the start and end position for each figurative unit, respectively. More precisely, FES-pan assigns each clause a binary tag (0/1), which indicates whether the current clause corresponds to a start or end position of a figurative unit.

4.2 Task 2: Figure Type Classification

Given a figurative unit comprising n words $\boldsymbol{u} = \{w_i\}_1^n$, and its context comprising m words $\mathcal{C} = \{w_j\}_1^m$, task 2 aims to classify this unit $\boldsymbol{u}$ into the right figure type.

First, the RoBERTa encoder produces contextualized representations h_i for each token w_i:

$$\boldsymbol{h} = (h_1, \ldots, h_n) = \text{Encoder}(w_1, \ldots, w_n) \quad (1)$$

Mean pooling is then applied on $\boldsymbol{h}$ to get hidden representation $h^{\boldsymbol{u}}$ for the figure unit $\boldsymbol{u}$. Based on this, we design the following baselines for task 2:

FTCLS FTCLS is built by adding a classification head on top of RoBERTa, which then classifies the input into one of 12 figure types.

FTCXT Since context information has proven to yield improvement for metaphor and sarcasm detection (Dankers et al., 2020; Mu et al., 2019; Joshi et al., 2016; Ghosh et al., 2017), FTCXT is designed to incorporate contextual information in figure classification and explore its effect across all figure types. Specifically, hidden representation $h^{\mathcal{C}}$ for context $\mathcal{C}$ is also calculated by the encoder. $h^{\boldsymbol{u}}$ and $h^{\mathcal{C}}$ are then concatenated for classification.

$$f = \text{Classification}([h^{\boldsymbol{u}}; h^{\mathcal{C}}]) \quad (2)$$

where f is the predicted figure type.

4.3 Task 3: Figure Recognition

Similar to task 1, given a discourse fragment $\mathcal{D}$ comprising m clauses and n words, Task 3 aims to extract figurative units $(u_1, \cdots, u_k)$ from $\mathcal{D}$, and classify each figurative unit into a specific figure type as $f_1, \cdots, f_k$ in the same time, where k is the number of figurative units in text $\mathcal{D}$.

Following task 1, contextualized representations $\boldsymbol{h_c} = (h_{c_1}, \ldots, h_{c_m})$ for the m clause are produced by RoBERTa encoder. Based on this, we define following baselines:

Rule-based Method Some figures in our dataset manifest obvious patterns. For example, the type of *Metaphor/Simile* usually comes with indicators such as the Chinese words "像 (like)", "如 (as)". The *Parallelism* type is composed of similar clauses, most commonly separated by colon.

To exploit these obvious patterns, we design heuristic rules for figure recognition as a complement to our neural methods.

Pipeline Since figure recognition can be naturally tackled as first extraction and then classification, we set our pipeline baseline as the combination of best-performing models in extracion task and classification task. Specifically, we first use FECRF to extract figurative units from input text, then use FTCXT to classify the extracted figurative units into their corresponding figure types.

E2ESeq By modeling figure recognition as a sequence labeling task, E2ESeq model shares the same architecture as the FESeq model mentioned before. The difference is that, in this case, we assign clauses of different figure types with different "B" and "I" labels. For instance, for a figurative unit of irony, we assign "B-Irony" to its first clause and "I-Irony" to other clauses in this unit.

E2ECRF Similar to FECRF, we add a CRF layer on top of the above E2ESeq model to construct a E2ECRF model.

5 Experiments

Implementation Details We implement our models using HuggingFace's Transformers (Wolf et al., 2020). We choose the RoBERTa-zh-Large (Cui et al., 2020) checkpoint trained on Chinese corpus. For fine-tuning, we generally stick to a dropout rate of 0.1, a batch size of 16, an epoch of 30, the Adam (Kingma and Ba, 2017) optimizer, and a learning rate of $1e^{-5}$. We select our hyperparameters based on the best performance on validation set and report the average results from 5 runs with different random seeds.

We present main experiment results on our dataset for all three tasks in Table 3. For the extraction task, we report precision, recall and Micro F1 score for each model. For other tasks, we additionally report Macro F1 score by averaging out F1 scores of all figure types. We also conduct subsidiary experiments and analysis for each task. By doing so, we shed light on promising directions for future work.

5.1 Evaluating Figure Extraction

Model performance for figure extraction is presented in the first block of Table 3. FECRF yields the best result, surpassing FESeq by 1.27 F1 score and FESpan by 1.77 F1 score, suggesting that figure extraction can be modeled as a sequence label-

Task	Model	Precision	Recall	Micro F1	Macro F1
Extraction	FESeq	34.06	29.23	31.46	-
	FECRF	32.56	**31.60**	**32.07**	-
	FESpan	**39.41**	24.61	30.30	-
Classification	FTCLS	78.82	78.82	78.82	65.22
	FTCXT	**79.49**	**79.49**	**79.49**	**68.72**
Recognition	Rule	7.74	12.76	9.64	5.24
	Pipeline	**31.10**	25.52	28.03	18.17
	E2ESeq	29.21	25.97	29.47	17.35
	E2ECRF	30.87	**31.10**	**30.99**	**21.27**

Table 3: Main experiment results. We highlight the highest numbers among models in **bold**.

Figure	Extraction (FESeq)	Recognition (E2ECRF)
Exact match	572	563
Wrong figure type	-	75
Super prediction	201	215
Sub prediction	289	227
Overlapping prediction	40	45
Non-lapping prediction	665	699
Total predictions	1757	1824

Table 4: Error analysis for model predictions in extraction and recognition task.

Figure	FTCLS			FTCXT		
	P	R	F1	P	R	F1
Meta.	84.16	87.12	85.61	83.40	88.29	85.78
Pers.	66.80	67.08	66.94	66.27	67.99	67.07
Meto.	62.24	49.59	55.20	66.67	47.15	55.24
Hyper.	57.66	54.48	56.03	58.06	49.66	53.53
Irony	50.00	20.00	28.57	50.00	40.00	44.44
Syna.	0.00	0.00	0.00	25.00	10.00	14.29
Rheq.	93.75	96.53	95.12	92.89	97.03	94.92
Para.	79.66	94.95	86.64	82.05	96.97	88.89
Dual.	70.13	71.05	70.59	70.13	71.05	70.59
Repe.	89.91	93.33	91.59	90.65	92.38	91.51
Anti.	67.31	56.45	61.40	78.43	64.52	70.80
Quote	85.34	84.62	84.98	87.93	87.18	87.55

Table 5: Figure type classification results w/ or w/o contextual information for each figure of speech. First seven figures are *tropes*, next five figures are *schemes*.

ing problem which CRF is specialized in. We also find that FESpan gives the highest precision but low recall value. This may be attributed to the fact that using two classifiers to predict the boundaries of figures imposes further restrictions.

Overall, the performance of figure unit extraction is not satisfactory, since it is somewhat difficult to precisely compartmentalize the smallest clause sequence containing figurative languages. Only 572 out 1757 predictions exactly matches gold label. We observe that the prediction errors can be largely categorized as follows: 1) *Super/Sub prediction*, the predicted clause sequence is a superset/subset of the gold figurative unit; 2) *Overlapping prediction*, the predicted clause sequence overlaps with a part of the gold figurative unit but is neither its superset nor subset; 3) *Non-lapping prediction*, the predicted clause sequence does not overlap with any gold figurative unit, i.e. predicting non-figurative clause groups as figurative. Column 2 in Table 4 presents the numbers of each error category according to prediction results of FESeq. More than half of the wrong predictions do not overlap with any gold figurative unit. Other wrong predictions are mainly supersets or subsets

of certain gold figurative unit. Based on these observations, we conclude that the SOTA models are not doing very well in either discerning figurative languages or determining the exact boundary of figurative unit in our corpus, perhaps because the size of our corpus is relatively small and imbalanced.

5.2 Evaluating Figure Type Classification

For figure type classification, we present results of FTCLS and FTCXT in the second block of Table 3. The latter gives an improvement of 0.67 micro F1 point and 3.5 macro F1 point over the former. Compared to figure extraction, the performance of figure type classification exhibits a high score.

Detailed classification results for each figure w/o contextual information is presented in Table 5. From this table, Columns 2-4 give FTCLS' classification results for each figure. It can be seen that classification accuracy is quite imbalanced across all figure types. On the one hand, F1 scores on *schemes* are relatively higher than *tropes* on the

whole. We suppose it is because *schemes* tend to manifest obvious patterns like repetition, which is easier for a model to capture, while *tropes* involve deviation from superficial meaning, which is more challenging to models and even humans. We expect more efforts on modeling semantics to benefit figure type classification. Notably, *Rhetoric Question (Rheq.)* gives the highest F1 score of 95.12 while being a trope. We suppose it is because the *question mark (i.e. "?") in Rheq.* units serve as an obvious clue to facilitate classification.

Table 5 reveals that even among trope figures, model performance is quite imbalanced. F1 score on Metahpor achieve 85.61 while on irony and synaesthesia it is 28.57 and 0.00, repectively. We suppose the reason for this performance gap is twofold. Firstly, figures such as irony and synaesthesia are demanding to collect, resulting in their fairly low distribution in ConFiguRe. To effectively train models for such low-frequency figures, it is necessary to incorporate other specialized techniques. Secondly, figures as epitomized by irony, are strongly related to a wider context, without which the identification of such figures becomes insufficient.

Context Benefits Figure Type Classification To inspect the extent that context information benefits figure classification, we include result of FTCXT in Columns 5-7 of Table 5. This serves as an "ablation study" that undergirds the usefulness of contextual information in figure classification. By comparing results of FTCLS and FTCXT, it can be seen that context information consistently improves classification accuracy for most figures. Further, among all figure types, contextual information is especially conducive to Irony, Synaesthesia and Antithesis, respectively boosting F1 score by 15.87, 14.29 and 9.40 point. This observation is consistent with the linguistic assumption that the identification of these figures usually depends on more context.

5.3 Evaluating Figure Recognition

For figure recognition, baseline results can be found in the third block of Table 3. Similar to figure extraction task, we observe that the sequence labeling model with CRF achieves the best result, surpassing the non-CRF version with 1.52 micro F1 score and 3.92 macro F1 score. It also outperforms the pipeline approach combining the best-performing extraction model and classification model, which suffers from error propagation. F1

| **Sequence 1:** 从胡乱对付事情这方面来说，你完全可以称得上是个天才（In terms of messing up with affairs, you may well be called a genius） |
| **Sequence 2:** 你完全可以称得上是个天才（You may well be called a genius） |

Figure 3: Example of two clause sequences marked as figurative unit for *irony*.

score of the rule-based method is the lowest compared to pipeline and end2end methods, indicating that figure recognition requires much more efforts beyond recognizing shallow and obvious patterns.

Same as figure extraction task, errors of this task incorporates the following types: *Super prediction*, *Sub prediction* and *Overlapping prediction*. Besides, we introduce *Wrong figure type* error, where the model successfully extracts the figurative units but classifies them into wrong figure type. According to the result of E2ECRF, numbers of each error type is presented in Column 3 of Table 4. We observe that a high percentage of precisely identified figure units are correctly classified with the figure type, and only 75 out of 638 predictions fail in classification. Resonating with figure extraction, we conclude that **discerning figurative units is quite challenging for state-of-the-art models.**

5.4 Discussion

Revisiting Figurative Unit From Subsection 5.1, we obtain a rather low performance in recognizing figurative units. We investigate the classification results of the models and find them rather satisfying. We therefore impute the poor performances to the difficulties in delimiting the boundary. Even if we clearly define a figure unit as the smallest clause sequence, such concept is somewhat controversial under certain circumstances. Figure 3 illustrates two instances which are ambiguous in deciding the gold figure unit. For this reason, it is necessary to improve the annotation process and evaluation metrics in the future.

Revisiting Tropes and Schemes Interestingly, our work can also serve as a supporting evidence for the linguistic categories of *tropes* and *schemes*. *Schemes* reflect deviation from the ordinary pattern or arrangement of words, while *tropes* involve deviation from the ordinary and principal significance of a word (Corbett, 1999). A heat map of confusion matrix in figure type classification is presented in Figure 4. It is intuitively suggested that, even misclassified, the predicted label tends
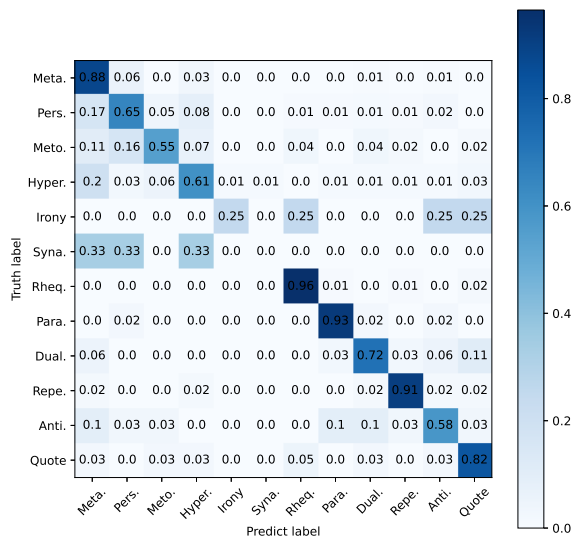
Figure 4: Confusion matrix from FTCLS's results in figure type classification. First seven figures are *tropes*, last five figures are *schemes*.

to fall into the same category as the gold label. For the example of *hyperbole*, there is a 61% probability of being classified correctly, a 20% probability as *metaphor/simile*, and a 6% probability as *metonymy*, all falling into the category of *tropes*. In the meantime, the probability of being classified as *schemes* is merely 7%. In other words, a trope figure is more similar to other trope figures than scheme figures, and vice versa.

6 Conclusions and Future Work

In this paper, we argue that it is necessary to recognize figures from the discourse level. For the first time, we propose the concept of figurative unit as language carrier of one figure and construct a Chinese corpus for Contextualized Figure Recognition (ConFiguRe). ConFiguRe includes 12 most commonly used figures in Chinese, with discourse-level contextual information attached to each figure instance. On ConFiguRe, we design three tasks of figure extraction, figure type classification and figure recognition, and implement state-of-the-art models. A series of experiments show that all three tasks remain challenging and worth exploring.

In future, we hope to increase the size of our corpus, especially adding more figure instances for the types with fewer instances such as irony and synaesthesia. At the same time, we will to improve the model performance with consideration of incorporating contextual information and using

more training data.

References

David Bamman and Noah Smith. 2015. Contextualized Sarcasm Detection on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):574–577. Number: 1.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 523–528, Valencia, Spain. Association for Computational Linguistics.

Gideon O. Burton. 2016. Silva rhetoricae.

Santiago Castro, Devamanyu Hazarika, Ver'şnica P'l'rez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards Multimodal Sarcasm Detection (An _obviously_ Perfect Paper). *arXiv:1906.01815 [cs]*. ArXiv: 1906.01815.

Xin Chen, Zhen Hai, Deyu Li, Suge Wang, and Dian Wang. 2021. Jointly Identifying Rhetoric and Implicit Emotions via Multi-Task Learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1429–1434, Online. Association for Computational Linguistics.

Edward Corbett. 1999. *Classical rhetoric for the modern student*. Oxford University Press, New York.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Verna Dankers, Karan Malhotra, Gaurav Kudva, Volodymyr Medentsiy, and Ekaterina Shutova. 2020. Being neighbourly: Neural metaphor identification in discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 227–234, Online. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter

and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.

Ellen K Dodge, Jisup Hong, and Elise Stickles. 2015. Metanet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49.

Paul Drew and Elizabeth Holt. 1998. Figures of speech: Figurative expressions and the management of topic transition in conversation. *Language in Society*, 27(4):495–522.

Dan Fass. 1991. Met*: A method for discriminating metonymy and metaphor by computer. *Comput. Linguist.*, 17(1):49–90.

Aniruddha Ghosh and Tony Veale. 2017. Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491, Copenhagen, Denmark. Association for Computational Linguistics.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The Role of Conversation Context for Sarcasm Detection in Online Interactions. *arXiv:1707.06226 [cs]*. ArXiv: 1707.06226.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A Report on the 2020 Sarcasm Detection Shared Task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online. Association for Computational Linguistics.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

The Pragglejaz Group, Raymond Gibbs, Alice Deignan, Graham Low, Gerard Steen, Lynne Cameron, Elena Semino, Joe Grady, Alan Cienki, and Zoltan Kövecses. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Randy Allen Harris, Chrysanne Di Marco, Sebastian Ruan, and Cliff O'Ă'ZReilly. 2018. An annotation scheme for Rhetorical Figures. *Argument & Computation*, 9(2):155–175.

Hyeju Jang, Seungwhan Moon, Yohan Jo, and Carolyn Rosé. 2015. Metaphor detection in discourse. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 384–392, Prague, Czech Republic. Association for Computational Linguistics.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic Sarcasm Detection: A Survey. *ACM Computing Surveys*, 50(5):73:1–73:22.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing Context Incongruity for Sarcasm Detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, and Mark J. Carman. 2016. Harnessing Sequence Labeling for Sarcasm Detection in Dialogue from TV Series 'Friends'. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 146–155, Berlin, Germany. Association for Computational Linguistics.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A large self-annotated corpus for sarcasm. *CoRR*, abs/1704.05579.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Maximilian Köper and Sabine Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30, Valencia, Spain. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282'Ă'Ş289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hankyol Lee, Youngjae Yu, and Gunhee Kim. 2020. Augmenting data for sarcasm detection with unlabeled conversation context. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 12–17, Online. Association for Computational Linguistics.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A Report on the 2018 VUA Metaphor Detection Shared Task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.

Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural Multitask Learning for Simile Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing, pages 1543–1553, Brussels, Belgium. Association for Computational Linguistics.

Jesse Mu, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Learning outside the box: Discourse-level features improve metaphor identification. 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 596–601, Minneapolis, Minnesota. Association for Computational Linguistics.

Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163.

Rossano Schifanella, Paloma de Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting Sarcasm in Multimodal Social Platforms. *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145. ArXiv: 1608.02289.

Steven Shapin. 2012. The ivory tower: the history of a figure of speech and its cultural uses. *The British Journal for the History of Science*, 45(1):1–27.

Gerard Steen. 2010. *A method for linguistic metaphor identification : from MIP to MIPVU*. John Benjamins Pub. Co, Amsterdam Philadelphia.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics.

Yi-jie Tang and Hsin-Hsi Chen. 2014. Chinese Irony Corpus Construction and Ironic Structure Analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1269–1278, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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, Baltimore, Maryland. Association for Computational Linguistics.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, Baltimore, Maryland. Association for Computational Linguistics.

Zhi Wen, Yang Li, Suge Wang, Jian Liao, and Xin Chen. 2019. Feature enhanced cnn for rhetorical questions identification. *Journal of Chinese information processing (in Chinese)*, 33(1):68–76.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, 6(1):53 – 74.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'l'mi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Gong Zhang. 1963. *Modern Chinese Rhetoric*. Tianjin People's Publishing House.

A Annotation Details

A.1 Figure Definition

We define these 12 figure categorites in reference to silva rhetoricae (http://rhetoric.byu.edu/), an authoritative website for rhetoric figures. We also follow a canonical work on rhetorics by Chinese linguistics Gong Zhang (1963). Detailed definition and example for each figure type above is provided in Figure 5.

A.2 Punctuation List

Below is punctuation list we use to separate clauses: { ，，；，。，？，！，……，——，：}

Figure	Definition	Example
比喻 (Metaphor/ Simile)	A comparison made by referring to one thing as another.	阳光像金子一样珍贵。(Sunshine is as precious as gold.)
比拟 (Personific ation)	Reference to abstractions or inanimate objects as though they had human qualities or abilities.	我心情愉悦，听见鸟儿在林中歌唱。(I was very happy and could hear the birds singing in the woods.)
借代 (Metonymy)	Reference to something or someone by naming one of its attributes.	据说他们是世界上最优秀的笔杆子。(They are said to be the best pens of the world.)
夸张 (Hyperbole)	Rhetorical exaggeration which is often accomplished via comparisons, similes, and metaphors.	他雄辩滔滔，仿佛能开岩裂石。(His eloquence would split rocks.)
反语(Irony)	Speaking in such a way as to imply the contrary of what one says, often for the purpose of derision, mockery, or jest.	这个勤奋的学生每周读书的时间从不超过一小时。(This diligent student seldom reads more than an hour per week.)
通感 (Synaesthe sia)	The production of a sense impression relating to one sense or part of the body by stimulation of another sense or part of the body.	品尝莫扎特音乐的味道。(Taste the music of Mozart.)
问语 (Rhetorical question)	Any question asked for a purpose other than to obtain the information the question asks.	这不是非常明显的例证吗？ (Isn't this a very obvious evidence?)
排比 (Parallelis m)	Similarity of structure in a pair or series of related words, phrases, or clauses.	童年是小草的芽儿，充满了生机；童年是早晨的太阳，充满了活力；童年是清润的雨水，充满了欢乐。(Childhood is the bud of grass, full of vitality; childhood is the morning sun, full of vitality; childhood is the rain, full of joy.)
对偶 (Duality)	Two similarly structured elements having the same length	不要对谦卑者傲慢，也不要对傲慢者谦卑。(Never be haughty to the humble; never be humble to the haughty.)
反复 (Repetition)	Repeat a word or a sentence.	他现在看起来比以前衰老很多很多了。(He looks much, much older now than before.)
对比 (Antithesis)	Juxtaposition of contrasting words or ideas (often, although not always, in parallel structure).	有的人活着，他已经死了；有的人死了，他还活着。(Some live, when they are already dead; others have died, but are still alive..)
引语 (Quote)	A group of words taken from a text or speech and repeated by someone other than the original author or speaker.	"众人拾柴火焰高"，让我们用我们自己的双手建设我们的家园！。(*The fire burns high when everybody adds wood to it*, let's build our gardon with our own hands.)

Figure 5: Definition and example for each figure type.