More Than Spoken Words: Nonverbal Message Extraction and Generation

Dian Yu Xiaoyang Wang Wanshun Chen Nan Du Longyue Wang
Haitao Mi Dong Yu
Tencent AI Lab
{yudian, shawnxywang, haitaomi, dyu}@global.tencent.com
{shaynechen, frankndu, vinnylywang}@tencent.com

Abstract

Nonverbal messages (NM) such as speakers' facial expressions and speed of speech are essential for face-to-face communication, and they can be regarded as implicit knowledge as they are usually not included in existing dialogue understanding or generation tasks. This paper introduces the task of extracting NMs in written text and generating NMs for spoken text. Previous studies merely focus on extracting NMs from relatively small-scale well-structured corpora such as movie scripts wherein NMs are enclosed in parentheses by scriptwriters, which greatly decreases the difficulty of extraction. To enable extracting NMs from unstructured corpora, we annotate the first NM extraction dataset for Chinese based on novels and develop three baselines to extract single-span or multi-span NM of a target utterance from its surrounding context. Furthermore, we use the extractors to extract 749K (context, utterance, NM) triples from Chinese novels and investigate whether we can use them to improve NM generation via semi-supervised learning. Experimental results demonstrate that the automatically extracted triples can serve as high-quality augmentation data of clean triples extracted from scripts to generate more relevant, fluent, valid, and factually consistent[1] NMs than the purely supervised generator, and the resulting generator can in turn help Chinese dialogue understanding tasks such as dialogue machine reading comprehension and emotion classification by simply adding the predicted "unspoken" NM to each utterance or narrative in inputs.

1 Introduction

Nonverbal messages (**NM**), such as facial expressions, body movements, and tones of voice, can complement or modify verbal messages as well as improve the teamwork efficiency (Breazeal et al., 2005) and effectiveness of face-to-face communication (Phutela, 2015). These messages are usually not explicitly mentioned in the transcribed verbal messages as the dialogue participants share most of the NMs via other modalities, and NMs are also seldom included in existing text-based dialogue tasks that mainly focus on verbal messages (Csaky and Recski, 2021). Though human readers can infer missing NMs based on their own knowledge, machines still have difficulty understanding the meanings behind and beyond the words (Zhang et al., 2018) and automatically decide what nonverbal behaviors they should display in interactions (Saunderson and Nejat, 2019). We focus on text-based NM extraction and generation, an important step towards reaching the ultimate goal of bridging the human-machine implicit knowledge gap.

One of the most relevant text resources for nonverbal messages is TV and movie scripts. Generally, scripts are written in a standard format: for example, NMs of their corresponding utterances are enclosed in parentheses (e.g., ELIZABETH (ironically) *"With five thousand a year, it would not matter if he had a big pink face."* and MR DARCY (shakes his head) *"You know how I detest it."*), which usually describe what can be seen or heard by the audience beyond the verbal messages. Based on the well-defined screenplay structures, it is relatively easy to use heuristics to extract utterances and their NMs from scripts (Vassiliou, 2006). However, in scripts, only a small percentage ($\approx 10.5\%$ (Section 6.2)) of utterances are followed by NMs, and existing public script corpora are usually small-scale even for resource-rich English (e.g., 1,276 movies (Gorinski and Lapata, 2015) and 917 movies (Gorinski and Lapata, 2018)).

In contrast, novels also contain rich NMs via the words of the writers alongside what their characters speak, and thousands of novels have already been adapted into scripts (mainly by professional scriptwriters). Besides, we observe that the density of NMs in novels is higher than that of scripts ($\approx 67.4\%$ based on the annotated corpus (Sec-

[1] See our metric definitions in Section 6.4 as different interpretations may exist for other generation tasks.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16396–16413
December 6-10, 2023

tion 3.2)), indicating the potential of leveraging novels for NM extraction. Therefore, we are interested in whether we can use this unstructured resource to alleviate the NM data scarcity problem, which hinders the full utilization of deep neural models. As this direction is still unexplored, we first define the task as extracting one or multiple spans from the surrounding context of the target utterance and annotate **NME**, the first **N**onverbal **M**essage **E**xtraction dataset based on three Chinese novels (Jia et al., 2021) containing 4K (context, utterance, NM) triples. Furthermore, we design three baselines (pattern, extractive, and generative) to extract NMs and evaluate them on NME.

Another question is whether we can leverage unlabeled novel corpora to automatically construct data for improving NM generation. To investigate this question, we first use the trained extractors to extract 749K pseudo-labeled triples from several hundreds of Chinese novels and train generators based on different backbone models to generate a nonverbal message given one target utterance and its context. Experiments show that these triples can serve as high-quality augmentation data of clean triples extracted from well-structured scripts to generate more relevant, fluent, valid, and factually consistent NMs. Furthermore, our semi-supervised generators can in turn help Chinese dialogue and narrative understanding tasks that lack NMs such as the dialogue subset of a machine reading comprehension dataset C^3 (Sun et al., 2020) and emotion classification EWECT by simply adding the generated "unspoken" NMs to each utterance or narrative in inputs, showing their usefulness.

The contributions of this paper are as follows.

- We design and annotate the first NM extraction dataset based on unstructured corpora.
- We design several strong nonverbal message extraction and generation baselines upon different backbone models to serve as a foundation for further work.[2]
- We extract large-scale (context, utterance, NM) data from unlabeled unstructured corpora using the NM extractors and demonstrate the usefulness of the data for improving the performance of NM generation.
- Experimental results show that NM generators can in turn help dialogue understanding tasks.

[2]We will release our code, annotated data, guidelines, parsed scripts, and models fine-tuned on the novel data at https://github.com/yudiandoris/nm.

2 Related Work

2.1 Nonverbal Message Extraction

Previous studies design patterns (Vassiliou, 2006; Wang, 2017; Sun et al., 2022) or language-specific features (Agarwal et al., 2014) to identify NMs, utterances, and speakers from well-structured (or semi-structured (Murtagh et al., 2009)) scripts in which NMs can be relatively easily extracted based on screenplay formats. To the best of our knowledge, this is the first work studying the automatic extraction of NMs from **unstructured** corpora.

2.2 Nonverbal Message Utilization

NMs such as facial expressions and body postures are used to facilitate tasks such as dialogue act classification (Ha et al., 2012), deception detection (Pérez-Rosas et al., 2015; Soldner et al., 2019), and text-to-speech (Kim et al., 2021). However, these messages are either pre-defined or converted based on recorded videos, making it resource-consuming to collect data and challenging to scale up for other applications. NMs extracted from scripts have been shown to be useful for tasks that require dialogue understanding such as machine reading comprehension and relation extraction (Sun et al., 2020). This work is the first attempt to use NMs extracted from **unstructured** data for **both understanding and generation tasks**.

3 NM Extraction Data Construction

As there is no public dataset in any language for identifying NMs in unstructured corpora, we first introduce our NM annotation guidelines and analysis of our annotated dataset NME.

3.1 Annotation Guidelines and Procedure

The annotation task is designed to select one or multiple spans (can be non-adjacent) from surrounding texts of a given utterance, as we aim to simplify the annotation task by copying the writers' words and avoid the annotators' interpretations that can be inaccurate. Since there can exist several NMs in the context, we only keep those that occur **a short time before** the utterance is spoken or **at the same time**, as they are more relevant to the target utterance (more examples in Appendix A.3). The speakers of utterances should not be included in the selected spans. In addition, we do not annotate relatively uninformative NM that only contains one or multiple of the three common speech verbs alone *"说"* and *"道"* (both means *"said"*) and

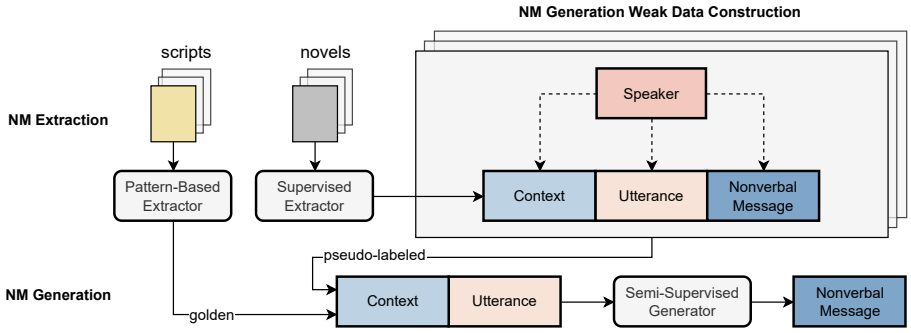

Figure 1: Overview of our framework (NM: nonverbal message. The supervised extractor is trained on NME).

"问" ("*asked*"). However, for consistent annotations, we ask the annotators to include these speech verbs when they appear right after other NMs such as "摇头道" ("*shook head and said*") and "笑道" ("*smiled and said*"). As an utterance may have a relatively uninformative NM or not be surrounded by any narrative texts, the annotated NM can be empty such as the last utterance "*Our enmity is as deep as the ocean...*" in Table 1.[3]

DOC	灭绝师太满脸怒容，说道"甚么明教？那是伤天害理，无恶不作的魔教。他……他躲在哪里？是在昆仑山的光明顶么？我这就找他去。" Miejue Shitai, with a furious expression on her face, said: "What Ming Cult? The Demon Cult is ferocious and inhuman, there are no evil deeds that they do not do. Where… where is he hiding now? Is he at Guang Ming Peak in the Kunlun Mountains? I'm going to look for him." 纪晓芙道："他说，他们明教……" Ji Xiaofu said: "He said he is the Ming Cult's…" 灭绝师太喝道："魔教！" Miejue Shitai shouted: "Demon Cult!" 纪晓芙道："是。他说，他们魔教的总坛，本来是在光明顶，但近年来他教中内部不和，他不便再住在光明顶，以免给人说他想当教主--→" Xiaofu said: "Yes. He said he is the Demon Cult's leader. Ordinarily, he would be at Guang Ming Peak but the last few years, there has been internal discord and fighting within the Cult. So he no longer lives on Guang Ming Peak to prevent people from thinking he wants to be the Cult Leader--→" 灭绝师太道："仇深似海！你大师伯孤鸿子，便是给这个大魔头杨逍活活气死的。"--→ Miejue Shitai said: "Our enmity is as deep as the ocean! Your Eldest Martial Uncle Guhong Zi was angered to death by the great demon Yang Xiao."--→
TU	"魔教！" ("Demon Cult!")
NM	喝道 (shouted)

Table 1: Example in the annotated nonverbal message extraction dataset NME (DOC: document. TU: target utterance. NM: nonverbal message. --→: omitted words)).

We construct unannotated NM data based on a public speaker identification dataset JY (Jia et al., 2021) in which three novels are involved. We simply use the annotated target utterance, its labeled speaker, and ten-sentence context (five sentences before/after the utterance) without any modifications. As "*speaker: utterance*" are regarded as two sentences separated by "*:*" in the JY dataset, the actual number of sentences in the context should be smaller than ten. Each triple (example in Table 1) is annotated by two annotators, and all annotators are Chinese native speakers. On average, each triple costs 0.30 RMB ($0.04) (more discussions in Ethical Considerations). For inter-annotation agreement (IAA) computation, when there are multiple annotated spans, we concatenate them into one span. IAA for span annotation measured by exact span match (Wang et al., 2021) is 83.0% and 92.0% by F1 (Hripcsak and Rothschild, 2005). The authors check and re-annotate the triples with different annotations for the final version.

3.2 Data Statistics and Analysis

The writers of novels tend to omit the NM when the utterance or the context is self-explanatory. Besides, triples with relatively uninformative NMs are discarded. Therefore, we only keep 67.4% of the annotated utterances, and each triple corresponds to one or multiple **non-empty** nonverbal messages (see discussions in Limitations). We report the data statistics in Table 2.

metric	value
number of training triples	3,791
number of development triples	341
average/max length (in characters) of context*	148 / 490
average/max length (in characters) of NM	4.7 / 33

Table 2: NME Statistics (*: exclude target utterance).

To examine the fine-grained types of NMs, we review the literature, analyze the annotated spans, and finally categorize them into thirteen sub-types (Table 3). See detailed definitions for each type and more examples in Appendix A.2. An NM may belong to multiple sub-types. For example, "陪笑" ("*put up a smiling face in order to please or pla-*

[3]English translations credit to https://wuxiasociety.com/, and they are not included in our dataset.

cate somebody") indicate both a facial expression and an intention, and *"均想"* (*"both think"*) shows both the number of speakers and the addressee. As shown in Figure 2, the two most frequent types are VOCAL-RELATED (e.g., pitch, volume, and speed) and KINESICS (i.e., body movement and facial expression), which are more expressive than other types such as INTENTION and therefore more likely to support downstream applications in other modalities such as speech and vision.

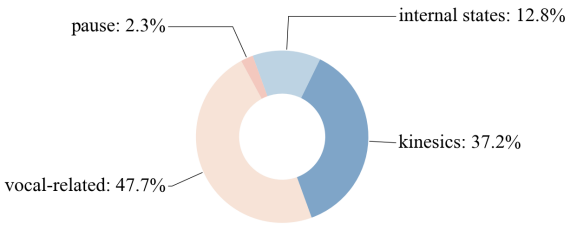

Figure 2: Distribution of nonverbal message types.

4 Nonverbal Message Extraction

This section introduces three NM extraction baselines (pattern-based, extractive, and generative) and how to use the extractors and unstructured corpora to construct large-scale pseudo-labeled NM data.

4.1 Pattern-Based Method

Based on the annotation guideline (Section 3.1), we can remove from considerations speaker names and utterances when we extract NMs. As the writers' own observations that may contain NMs take place alternating with utterances of characters (Poyatos, 1977), we assume that an NM is very likely to appear in the same paragraph as the given target utterance. We first run a strong ($\approx$ 90% in F1) extractive speaker identification model (Yu et al., 2022) over the paragraph to identify the speaker of the target utterance. As utterance annotations are unavailable in unlabeled novels, we use double quotation marks to segment utterances and regard the first one as the target utterance. Then we remove all utterances and the speaker from the paragraph, separated the remaining context by commas, and use the last span as the NM of the target utterance to reduce noise. For example, given a paragraph *"Miejue Shitai shouted 'Demon Cult!'"*, our pattern-based extractor extracts *"shouted"* as the NM of the underlined utterance. However, this method will inevitably suffer from relatively low recall (e.g., 63.1% in macro-averaged recall on the dev set of NME). For example, given the context of E6 in Table 3, both *"heart was relieved"* and *"thought"* should be regarded as NMs while this method can only extract *"thought"*.

4.2 Extractive Method

As a nonverbal message mention must be a one or multiple spans in the context surrounding the target utterance, we consider an extractive machine reading comprehension (MRC) formulation (Devlin et al., 2019) that originally aims to extract an answer of a given question from a document.

We regard the target utterance u as the question and regard the surrounding context of u as well as u as document d. The ground truth nonverbal message of u is treated as the answer a. We follow previous work to concatenate a special token [CLS], tokens in u, a special token [SEP], and tokens in d as the input sequence. Two vectors p_{start} and p_{end} are introduced to represent the estimated probabilities of each token in d to be the start or end token of the correct answer span a that appears in d, respectively. Let a_{start} and a_{end} denote the start offset and end offset of a, respectively. We optimize the extractive model with parameters θ by minimizing $\sum_{t \in V} L(t, \theta)$, where V represents the set of NM extraction instances, and L is defined as:

$$L(t,\theta) = -\log p_{\text{start},\theta}(a_{\text{start}} \mid t) - \log p_{\text{end},\theta}(a_{\text{end}} \mid t). \quad (1)$$

However, this classical extractive architecture can only extract a **single** span from the context, though there are some attempts (e.g., (Segal et al., 2020; Yang et al., 2020)) to extend it to extract multiple spans. For those instances with multiple NMs, we simply use the longest common substring of the context and the concatenation of these NMs as the actual answers for training and the original labels for validation. We introduce multi-span formulation in the following subsection.

4.3 Generative Method

To address the single-span limitation of the above extractive method, we regard nonverbal message extraction as a text-to-text task (Raffel et al., 2020): the extractor is fed the surrounding context of the target utterance and is asked to generate the NM of this utterance. For NMs that are a set of non-contiguous spans, we concatenate them using commas to form the ground truth labels.

Type	Sub-Type	Example	ID
kinesics	body movement	⊢双手紧紧抓住被角⊣。颤声道⊣: “他...他...怎么了？ ” ⊢grabbed the corner of her quilt tightly with both hands⊣ and asked in a shaking voice⊣, “He... What... What happened to him?”	E1
	facial expression	..., 当即⊢嘿嘿冷笑⊣，说道: “你要命不要...” ... he ⊢sneered⊣ and said, “Are you bored of your life? ...”	E2
internal states	intention	朱九真道: “...” 话声中带着三分小女孩儿的⊢撒娇⊣意。 Zhu Jiuzhen said, “...” Her voice carried ⊢a thirty-percent little spoiled girl's tone⊣.	E3
	feelings/emotions	..., 随即⊢心下欣喜⊣，心道⊣: ... ... immediately ⊢his heart was relieved⊣, and he thought⊣ ...	E4
pause	—	⊢征了片刻⊣，便道: “多蒙前辈手下留情。” ⊢After a moment⊣, he said, “Thank you for not taking my life.”	E5
vocal-related	addressee	..., 随即⊢心下欣喜⊣, ⊢心道⊣: ... ... immediately ⊢his heart was relieved⊣ and he ⊢thought⊣ ...	E6
	# speakers	朱元璋、徐达、常遇春等⊢齐道⊣: ... Zhu Yuanzhang, Xu Da and the others said ⊢in unison⊣ ...	E7
	tone	何足道⊢怒道⊣: “少林寺卧虎藏龙之地...” He ZuDao ⊢angrily said⊣: “The ShaoLin Monastery has indeed many extraordinary people ...”	E8
	volume	赵敏⊢低声道⊣: “最好有人...” ⊢In a low voice⊣ Zhao Min said, “It would be better if someone ...”	E9
	speed	... 见杨过脸色沉重，只为自己担忧，⊢缓缓的道⊣: ... Seeing a melancholy expression on his face, she ⊢slowly said⊣ ...	E10
	pitch	那身矮老者⊢尖声说道⊣: “姓曾的...” the short old man ⊢says sharply⊣, “Mr. Zeng ...”	E11
	timbre	只听得左边旗斗中一个⊢苍老的声音⊣哈哈大笑⊣，说道: ... ⊢An old voice⊣ from atop one of the flag poles ⊢laughed⊣...	E12
	others	⊢双手紧紧抓住被角⊣。颤声道⊣: “他...他...怎么了？ ” ... ⊢grabbed the corner of her quilt tightly with both hands and asked⊣ in a shaking voice⊣, “He... What... What happened to him?”	E13

Table 3: Examples of nonverbal messages in novels (⊢ (⊣): start (end) position of one annotated nonverbal message).

We follow previous work to minimize a corruption objective $\sum_{t \in T} L(t, \theta_e, \theta_d)$ to train an encoder-decoder model over data T:

$$L(t, \theta_e, \theta_d) = -\log p_{\theta_e, \theta_d}(y_t \mid x_t), \tag{2}$$

where output y_t is the NM of the given utterance in instance t, input x_t is the surrounding context of this utterance, and θ_e and θ_d represent the parameters for the encoder and decoder, respectively. In the input sequence, the preceding/following context is separated by special tokens [SEP]. We experiment with different types of input (Table 4). Despite its advantage of extracting multi-span messages, compared with the faithful NMs extracted by the pattern-based and extractive methods, generative methods may have hallucination issues. See more discussions in Section 6.3.

4.4 Large-Scale Weak Data Construction

We use the pattern-based extractor and the speaker identification model (Section 4.1) to extract (context, utterance, NM) candidates with non-empty NMs from hundreds of novels. For data quality control, we discard those instances whose NMs have more than nine characters or are relatively uninformative (defined in Section 3.1). Note if the paragraph before the target utterance also contains utterances, we use their speaker and the last utterance (separated by a colon) as the context to make the data format consistent with that of scripts. If no utterance exists in the previous paragraph, all the content of the paragraph is used as context. To obtain the NMs extracted by the trained extractive and generative extractors, which are used to replace pattern-based NMs in (context, utterance, NM) triples, we use the paragraph that includes the target utterance and its previous paragraph as context for inference. In summary, we construct the same number of (context, utterance, NM) triples using each of the three extraction methods introduced in this section to extract corresponding NMs.

5 Nonverbal Message Generation

5.1 Method

Given a history context sequence that contains k utterances or narratives $U = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$, the task aims to generate a natural language nonverbal message $\mathbf{n} = \{n_1, n_2, \ldots, n_m\}$ for speaker of the k-th utterance $\mathbf{u}_k$, where m denotes the maximum possible number of words in the NM. The probability of the NM is formulated as:

$$p(\mathbf{n} \mid U) = \prod_{i=1}^{m} p(n_i \mid U, n_1, \ldots, n_{i-1}) \tag{3}$$

Similar to the generative extraction method (Section 4.3), we adopt the text-to-text formulation and train the generator with a maximum likelihood to predict the target sequence.

5.2 Training Paradigm with Automatically Constructed Data

To leverage the weakly-labeled nonverbal message generation instances constructed by extraction methods introduced in Section 4.4, we conduct

two-stage fine-tuning, which has been widely used in previous semi-supervised studies to reduce the impact of noise (Xie et al., 2020; Sun et al., 2022): we first fine-tune a generator on the combination of the clean and weak-labeled data and then fine-tune the resulting generator on the clean data alone.

6 Experiment

6.1 Implementations

For the extractive baselines, we consider both encoder-only (RoBERTa-wwm-ext-large and MacBERT, both released by Cui et al. (2021)) and encoder-decoder models ($T5_{base}$ (Zhao et al., 2019), $BART_{large}$ (Shao et al., 2021), and $DialBART_{large}$). Note that $DialBART_{large}$ (HIT-TMG, 2022) is obtained by fine-tuning $BART_{large}$ (Shao et al., 2021) on $LUGE_{dialogue}$ (Section 6.2). The above encoder-decoder models are also used for generative NM extractors and NM generators. We discard the tokens from the bottom of the input if its length exceeds the maximum model sequence length. We conduct experiments on eight NVIDIA-V100 32GB GPUs. Appendix A.8 introduces the models' details (Table 19), hyper-parameters (Table 18) for each task, and evaluation metrics. We run each experiment five times with different random seeds.

6.2 Datasets

Script: we use Chinese scripts to construct clean NM generation data. We collect scripts from a script website[4] ONLY available for research and non-commercial use and keep 454 scripts after filtering those with format issues. We introduce more details about how to parse scripts in Appendix A.1 as this is not this work's main contribution and has been widely explored in the literature. We use triples extracted from non-overlapped scripts as training and dev sets to avoid data leakage. The most recent 50 scripts are used for the dev set.

Novel: we collect 521 Chinese novels from Yuewen for weak-labeled NM generation data construction (Section 4.4). Due to copyright restrictions, the novels will not be directly released. We experiment with two weak NM generation data of different sizes: Novel (397K) and $Novel_L$ (749K), and Novel is a subset of $Novel_L$.

Commonsense Knowledge (CSK): a context-utterance-NM triple can somehow be regarded as a piece of commonsense knowledge, as they both are usually ignored and assumed to be known without being told. Thus, we use the human-annotated argument pairs such as (*"someone holds an umbrella"*, *"it is raining"*) in the Chinese set of commonsense knowledge ConceptNet (Speer et al., 2017) by regarding the two arguments as the input and the NM, respectively. The context is left empty.

$LUGE_{dialogue}$: it contains four Chinese dialogue datasets: Chinese Persona Chat (CPC)[5], LCCC (Wang et al., 2020), Emotional STC (ESTC) (Zhou et al., 2018), and KdConv (Zhou et al., 2020). We indirectly use $LUGE_{dialogue}$ by using $DialBART_{large}$ that is fine-tuned on this dataset as the backbone model to study the usefulness of dialogue generation datasets for NM generation.

C^3_D and **EWECT**[6]: for dialogue/narrative understanding tasks, we consider a multiple-choice MRC dataset C^3_D and an emotion classification dataset EWECT (the general-domain version) to investigate the impact of introducing generated NMs into dialogue tasks without NMs. We provide the detailed data statistics in Table 12 (Appendix A.4).

6.3 NM Extraction Evaluation

Supervised extractors outperform the pattern-based method (Table 4). We find that including the target utterance in the input hurts the performance of generative methods even though the utterance boundary is indicated by [SEP]. Methods such as increasing the training data size (clean or noisy) of NME may help models learn to focus on the writers' words for identifying NMs. The length distribution of the NMs extracted by our supervised extractors is very similar to that of the clean NMs in Script. Note that the sharp drop in the pattern-based NM distribution is caused by the length constraint we set for weak data construction (Section 4.4). We also find that the generative models underperform in instances with multiple NMs (Appendix A.3).

6.4 NM Generation Evaluation

For the majority baseline, we use the most frequent NM (1.03%), *"笑"* (*"smile"*), in the training set of the Script as the NMs for all utterances. We notice a model pre-fine-tuned on dialogue generation datasets $LUGE_{dialogue}$ performs better on NM generation (6 vs. 4 in Table 5). For semi-supervised training, we experiment with two backbone models ($T5_{base}$ and $DialBART_{large}$) and see consistent gains (10 vs. 3) (15/18 vs. 6) over the purely super-

[4]https://www.1bianju.com.

[5]https://www.luge.ai.

[6]https://smp2020ewect.github.io/.

extractor	model	input	EM↑	F1↑
pattern-based	–	cat(context_1, context_2)	41.6	62.5
extractive	$\text{MacBERT}_{\text{large}}$	utterance [SEP] cat(context_1, utterance, context_2)	74.3 (0.81)	88.1 (0.41)
extractive	RoBERTa-wwm-ext-large	utterance [SEP] cat(context_1, utterance, context_2)	74.9 (0.33)	88.5 (0.21)
generative	$\text{BART}_{\text{large}}$	context_1	72.0 (3.24)	86.7 (0.66)
generative	T5_{base}	context_1	74.0 (1.11)	88.0 (0.37)
generative	$\text{DialBART}_{\text{large}}$	context_1	**76.4** (0.79)	**89.0** (0.39)
generative	$\text{DialBART}_{\text{large}}$	context_1 [SEP] context_2	75.0 (0.74)	88.1 (0.60)
generative	$\text{DialBART}_{\text{large}}$	context_1 [SEP] utterance [SEP] context_2	74.2 (0.66)	87.9 (0.62)

Table 4: The nonverbal message extraction performance on the NME dataset (cat: concatenation; context_1: context before the target utterance. context_2: context after the target utterance. EM: exact match).

training data/method	model	type	F1↑	ROUGE-1↑	ROUGE-L↑	DIST-1↑	DIST-2↑	ID
majority	–	weak	3.18	3.20	3.20	0.06	0.06	1
CSK (Speer et al., 2017)	T5_{base}	indirect	1.94 (0.15)	2.02 (0.15)	1.99 (0.16)	10.05 (2.17)	17.44 (2.61)	2
$\text{Script}_{\text{train}}$	T5_{base}	direct	7.20 (1.26)	7.73 (1.35)	7.62 (1.33)	7.47 (4.07)	15.32 (8.42)	3
$\text{Script}_{\text{train}}$	$\text{BART}_{\text{large}}$	direct	7.49 (0.56)	8.05 (0.74)	7.92 (0.73)	6.65 (3.52)	13.22 (7.17)	4
–	$\text{DialBART}_{\text{large}}$	indirect*	2.92	3.36	3.23	10.22	43.35	5
$\text{Script}_{\text{train}}$	$\text{DialBART}_{\text{large}}$	semi*	7.65 (0.31)	8.29 (0.30)	8.15 (0.32)	10.12 (0.40)	20.03 (0.68)	6
$\text{Novel}_{\text{pattern}}$	T5_{base}	weak	4.58 (0.18)	4.96 (0.18)	4.89 (0.18)	5.51 (0.28)	11.22 (0.63)	7
$\text{Novel}_{\text{extractive}}$	T5_{base}	weak	4.54 (0.18)	4.95 (0.21)	4.92 (0.21)	14.68 (0.28)	36.69 (0.94)	8
$\text{Novel}_{\text{generative}}$	T5_{base}	weak	4.83 (0.23)	5.28 (0.26)	5.24 (0.26)	15.64 (0.17)	40.01 (0.82)	9
$\text{Novel}_{\text{pattern}} \twoheadrightarrow \text{Script}_{\text{train}}$	T5_{base}	semi	**7.85** (0.23)	**8.59** (0.23)	**8.46** (0.25)	9.93 (0.35)	22.32 (0.62)	10
$\text{Novel}_{\text{extractive}} \twoheadrightarrow \text{Script}_{\text{train}}$	T5_{base}	semi	7.53 (0.25)	8.18 (0.26)	8.08 (0.28)	10.01 (0.25)	22.32 (0.82)	11
$\text{Novel}_{\text{generative}} \twoheadrightarrow \text{Script}_{\text{train}}$	T5_{base}	semi	7.83 (0.37)	8.44 (0.37)	8.33 (0.35)	10.37 (0.40)	22.76 (0.54)	12
$\text{Novel}_{\text{pattern}} \twoheadrightarrow \text{Script}_{\text{train}}$	$\text{DialBART}_{\text{large}}$	semi	7.87 (0.25)	8.50 (0.28)	8.40 (0.27)	10.15 (0.62)	21.15 (1.43)	13
$\text{Novel}_{\text{extractive}} \twoheadrightarrow \text{Script}_{\text{train}}$	$\text{DialBART}_{\text{large}}$	semi	7.77 (0.17)	8.44 (0.21)	8.34 (0.17)	10.44 (0.46)	21.28 (0.66)	14
$\text{Novel}_{\text{generative}} \twoheadrightarrow \text{Script}_{\text{train}}$	$\text{DialBART}_{\text{large}}$	semi	**7.95** (0.22)	**8.68** (0.25)	**8.56** (0.24)	10.45 (0.42)	21.80 (0.95)	15
$\text{Novel}_{\text{pattern}}$ (L) $\twoheadrightarrow \text{Script}_{\text{train}}$	$\text{DialBART}_{\text{large}}$	semi	8.05 (0.24)	8.71 (0.25)	8.60 (0.25)	10.02 (0.36)	20.90 (0.78)	16
$\text{Novel}_{\text{extractive}}$ (L) $\twoheadrightarrow \text{Script}_{\text{train}}$	$\text{DialBART}_{\text{large}}$	semi	7.87 (0.18)	8.55 (0.16)	8.44 (0.18)	10.62 (0.42)	21.39 (0.92)	17
$\text{Novel}_{\text{generative}}$ (L) $\twoheadrightarrow \text{Script}_{\text{train}}$	$\text{DialBART}_{\text{large}}$	semi	**8.15** (0.28)	**8.84** (0.29)	**8.72** (0.28)	10.22 (0.30)	21.64 (0.99)	18

Table 5: The nonverbal generation average performance and standard deviation on the dev set of the Script ($\twoheadrightarrow$: two-stage fine-tuning (Section 5.2). $\star$: as $\text{DialBART}_{\text{large}}$ is pre-fine-tuned on $\text{LUGE}_{\text{dialogue}}$).

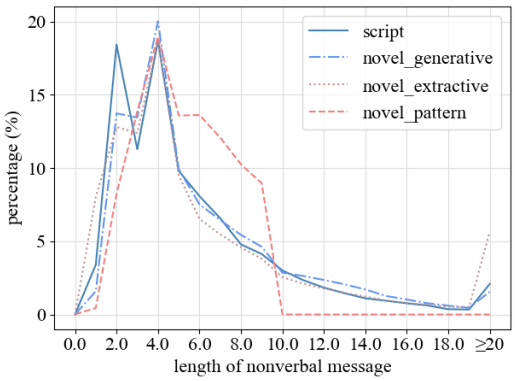

Figure 3: Length distribution of nonverbal messages extracted by scripts and our three methods.

input	F1↑	ROUGE-1↑	DIST-2↑
u_{k-1}, u_k	4.83 (0.23)	5.28 (0.26)	40.01 (0.82)
u_k	4.77 (0.23)	5.05 (0.24)	6.11 (0.43)
u_{k-2}, u_{k-1}, u_k	4.78 (0.22)	5.24 (0.23)	38.50 (1.11)
u_{k-1}, s_k, u_k	4.85 (0.23)	5.28 (0.25)	41.98 (0.96)

Table 6: The impact of introducing context and speaker information for semi-supervised training based on the automatically constructed data ($\text{Novel}_{\text{generative}}$) ($u_k$: the k-th utterance, i.e., the target utterance, s_k: the speaker of the target utterance; different components in the input are separated by [SEP]).

vised baselines trained on Script. Introducing more weakly-labeled data is also helpful (18 vs. 15).

Although the extractive extractor outperforms the pattern-based one by a large margin on the annotated nonverbal message extraction dataset (74.9% vs. 41.6% in Table 4), this supervised baseline does not outperform the pattern-based one when used to construct weakly-labeled nonverbal message generation data. One possible reason for this could be that the extractive baseline is more likely to predict a wrong span boundary by including irrelevant long context in NMs compared with other baselines. This is also commonly seen when this formulation is used for other span extraction tasks (Gao et al., 2019). As shown in Figure 3, more than 5% of the NMs extracted from novels by the extractive baseline contain more than 20 characters, and the length inconsistency of the resulting data may hinder the training of NM generators.

The Impact of Context and Speaker ID: to investigate the impact of context on NM generation, remove the context from the training instances. In

metrics	input sequence (context [SEP] target utterance)	prediction	gold label
relevance	女店员:请问是要结婚戒指吗? [SEP] 不是。看这个吧 Shopwoman: Are you looking for a weding ring? [SEP] No. I would like to see this one.	拿出戒指 takes out the ring	看中一款名贵的戒指 interested in an expensive ring
fluency	石浩:爸爸，我饿了。[SEP] 好，快吃吧，快吃饭吧，尝尝这个，这个，是我老婆亲手腌的，特别脆，辣味会一下子冲进来，...., 尝尝看，尝尝看～ Shi Hao: Dad, I'm hungry. [SEP] Okay, let's eat! Try this. This is marinated by my wife. It's very crispy, and the spicy taste will burst in all at once,..... Try it!	拿出一个小包 take out a small bag (bun?)	手势形容 describes with gestures
validity	万宗华:沉肩，坠肘，抱圆。瞧瞧你练的是什么啊。就这样，怎么参加中秋晚会？怎么表演？ [SEP] God，this is so tedious.* Wan Zonghua: Sink shoulders, drop elbows, and embrace circles. Look what you are practicing. Just like that, how can I participate in the Mid-Autumn Festival Gala? How to perform? [SEP]God，this is so tedious.	看着舞台上的舞台 looks at the stage on the stage	小声嘀咕 whispers
factual consistency	两人走在深夜的森林中。[SEP] 伤还没好，要不要试试我的独家秘方? Two people were walking in the woods late at night. [SEP] The injury is still not healed, do you want to try my exclusive secret recipe?	看着她的手 looks at her hand	拿出一条小鱼 takes out a small fish

Table 7: Some negative examples for each metric (one may belong to multiple types) (*originally in English).

other words, input becomes the target utterance alone. We use $T5_{base}$ for the ablation studies. Based on the input sequence we use in the main experiments, we additionally add extra preceding context (i.e., one narrative or utterance) or the speaker(s) of the target utterance u_k. As shown in Table 6, the context before the target utterance is important for NM generation, while introducing more history context (u_{k-2}) or speaker(s) of u_k does not lead to notable performance improvement. Also, the result indicates that other dialogue-related datasets without speaker information may also be considered to be used for improving NM generation.

Human Evaluation and Error Analysis: we randomly sample 100 instances from the held-out set of Script and randomly shuffle the label and automatically generated NMs for each instance. Given the context, target utterance, and an NM, we ask annotators to rate each NM using the following four binary metrics: (M1) the relevance between the utterance and the NM based on the context, (M2) the fluency of the NM, (M3) the validity of the NM, and (M4) the factual consistency of the NM based on the context and utterance. Table 7 shows some negative examples for each metric. We elaborate on our evaluation guidelines and provide error analysis showing models may need more external knowledge from different resources in Appendix A.5.

For the NM generation human evaluation, the human agreement (κ) is measured using Cohen's kappa (details in Appendix A.5). For all four metrics, $\kappa = 0.55$ (moderate agreement). When we do not consider the hallucination issue in M4 as ground truth NM label cannot be judged using this metric, $\kappa = 0.64$ (substantial agreement). Similar to our observations when automatic metrics are used, models trained with automatically extracted data achieve better performance over the purely supervised baseline trained with Script (Table 8).

description	M1↑	M2↑	M3↑	M4↑	AVG4	AVG3
$\text{Script}_{\text{train}}$	61.3	70.0	76.0	64.7	68.0	69.1
$\text{Novel}_{\text{pattern}}$ (L)	**66.0**	73.3	77.3	81.3	74.5	72.2
$\text{Novel}_{\text{extractive}}$ (L)	50.7	76.0	85.3	88.0	75.0	70.7
$\text{Novel}_{\text{generative}}$ (L)	62.7	**81.3**	**86.0**	**88.0**	**79.5**	**76.7**
ground truth	72.0	77.3	83.3	–	–	77.6

Table 8: Human evaluation (%) on the held-out set of Script (M1: relevance, M2: fluency, M3: validity, M4: consistency, AVG4/3: average of M1–4 and M1–3).

6.5 Evaluation on Natural Language Understanding Tasks

We also study whether the predicted NMs (by different generators in Table 5) can in turn help dialogue/narrative understanding tasks. We adopt the baselines released by Zhao et al. (2019). For each utterance in the input of each instance, we add a predicted NM after the utterance and keep other settings unchanged. For example, for C^3_D that aims to select the correct answer option of a question based on a dialogue, one modified dialogue example (English translation) is *"Female: Hey, where are you? We are all waiting for you! (hurriedly shouting) Male: Immediately, I'll be there soon! I've already got off the bus, and I'm on my way to you! (hurriedly said)"*. We conduct the same procedure for EWECT that aims to identify the emotional state of a writer or speaker. We see significant improvements by introducing the NMs into the original tasks (Table 9) in a human-interpretable way compared with the same implemented baseline without considering NMs, and NMs have a similar impact

as the clean commonsense and script knowledge on the two tasks. In Appendix A.7, we show complete expanded examples for each task in Table 16 and provide error analysis for future directions.

external source	backbone	# mpu*	C^3_D dev	C^3_D test
–	$\text{BERT}_{\text{BASE}}$	0	62.3§	62.1§
–	$\text{RoBERTa}_{\text{LARGE}}$	0	67.3 (3.45)	66.1 (2.19)
CSK	$\text{RoBERTa}_{\text{LARGE}}$	1	68.3 (0.72)	67.0 (0.89)
Script	$\text{RoBERTa}_{\text{LARGE}}$	1	68.8 (1.19)	66.3 (1.03)
Novel_L	$\text{RoBERTa}_{\text{LARGE}}$	1	**69.4** (0.46)‡	**67.6** (0.89)‡
external source	backbone	# mpu*	EWECT dev	EWECT test
–	$\text{BERT}_{\text{LARGE}}$	0	78.7§	–
–	$\text{RoBERTa}_{\text{LARGE}}$	0	79.7 (0.25)	78.4 (0.34)
CSK	$\text{RoBERTa}_{\text{LARGE}}$	1	79.4 (0.49)	78.9 (0.20)
Script	$\text{RoBERTa}_{\text{LARGE}}$	1	79.4 (0.31)	**79.0** (0.47)
Novel_L	$\text{RoBERTa}_{\text{LARGE}}$	1	**80.2** (0.19)†	78.8 (0.24)‡

Table 9: The accuracy (%) of introducing NMs into C^3_D and EWECT (mpu*: number of added NMs per utterance in the original input. $\text{RoBERTa}_{\text{LARGE}}$: RoBERTa-wwm-ext-large. §: copied from (Sun et al., 2020) and top-1 team's report from the EWECT website. ‡: p-value < 0.005. †: p-value < 0.05).

7 Conclusions and Future Work

This work focuses on NM extraction and generation leveraging unlabeled corpora, and experimental results show the effectiveness of our proposed methods and the usefulness of the large-scale weakly-labeled data. Future work includes improving NM extraction via semi-supervised learning, incorporating speaker profiles into NM generation, and generating both utterances and their NMs (Section A.6).

Acknowledgements

We would like to thank the anonymous meta-reviewers/reviewers for their insightful feedback.

Limitations

Scope and Generalizability: though code-switch may occur in utterances, all the datasets (NME, weak-labeled NM data, and downstream tasks) are mainly written in Chinese. Though our methods especially the supervised ones do not consider many Chinese-specific features, it is unclear to what extent the differences in grammar and culture across different languages or even within one language will impact the generalizability of empirical results and observations based on data in a single language used in this paper. In particular, it has been revealed by nonverbal communication literature (LaFrance and Mayo, 1978) that considerable culture-related differences are more likely to exist for interpersonal NMs: for example, bowing is used for showing status differences in some cultures such as Japan (Morsbach, 1973). In addition, NME is based on an existing novel dataset that involves three novels of similar genres written by a single author. Though this may decrease the difficulty of NM annotation, the diversity of NMs is relatively insufficient, which may hurt the robustness and generalization abilities of the supervised NM extractor trained on NME.

Silent NM: it is possible that an utterance is spoken without any CLEAR nonverbal messages. This work only focuses on utterances with non-empty nonverbal messages when we build clean NM extraction data (Section 3.1), and we use the pattern-based extractor as the first step of building large-scale weakly-labeled data (Section 4.4), which also helps filter instances with empty NMs.

To identify silent NMs, besides the pattern-based method in this work, the discarded instances with empty NM can be kept to train an extractor (e.g., similar to the formulation and implementation to address the unanswerable questions in machine reading comprehension tasks (Rajpurkar et al., 2018)) or generator (e.g., simply using a predefined answer such as *"empty"* or *"none"* as the NM of these instances), though this will inevitably increase the task difficulty.

Context Selection: for our NM generation experiments except for the ablation studies, we train and evaluate our models on instances with one history utterance or narrative. As the target utterance is more important for NM generation than dialogue history, this does not lead to a negative impact on the performance. However, introducing more context can facilitate annotators' understanding to better score a generated or ground truth NM, and this is necessary if we aim to generate diverse target utterances as well as their NMs.

Human Evaluation: for each criterion, we simply ask annotators to rate 0 or 1 for a nonverbal message, as NM is shorter (e.g., a verb phrase) than the utterance outputs of dialogue generation tasks that may allow more fine-grained scales. We admit that the evaluation guidelines can be further improved.

Ethical Considerations

Based on the authors' experience, the price is cheaper than that of similar span selection tasks in earlier years, perhaps due to the impact of the epidemic and supply and demand in China when the annotation job was submitted in 2022. The NM extraction annotation is conducted by annotators from a commercial annotation company (large enterprise[7]) headquartered in Chengdu, Sichuan.[8] A four-day trial annotation is conducted for bidding (seven companies participated), and the authors answered questions posted by the annotators from different companies and updated annotation guidelines accordingly. The Q&A history is maintained and updated in the formal annotation. After the pilot annotation, the quoted price from the selected company is $0.15 \times 2 = 0.30$ RMB ($0.04) for each instance. The formal annotation including data acceptance is finished within one week.

References

Apoorv Agarwal, Sriramkumar Balasubramanian, Jiehan Zheng, and Sarthak Dash. 2014. Parsing screenplays for extracting social networks from movies. In *Proceedings of the CLFL*, pages 50–58.

Nalini Ambady and Robert Rosenthal. 1998. Nonverbal communication. *Encyclopedia of mental health*, 2:775–782.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Proceedings of the IROS*, pages 708–713.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the NeurIPS*, pages 1877–1901.

Michela Cecot. 2001. Pauses in simultaneous interpretation: A contrastive analysis of professional interpreters' performances. *The interpreters' newsletter*, 11:63–85.

[7] based on the information provided by https://www.kanzhun.com/firm.

[8] http://sjbz.itaojin.cn.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint*, cs.CL/2006.14799v2.

Herbert H Clark and Thomas B Carlson. 1982. Hearers and speech acts. *Language*, 58(2):332–373.

Richard Csaky and Gábor Recski. 2021. The Gutenberg dialogue dataset. In *Proceedings of the EACL*, pages 138–159.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *TASLP*, 29:3504–3514.

Joseph A DeVito, Susan O'Rourke, and Linda O'Neill. 2000. *Human communication*. Longman New York.

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 NAACL-HLT*, pages 4171–4186.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialog state tracking: A neural reading comprehension approach. In *Proceedings of the SIGDIAL*, pages 264–273.

Daniel Gatica-Perez. 2009. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and vision computing*, 27(12):1775–1787.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the NAACL-HLT*, pages 1066–1076.

Philip John Gorinski and Mirella Lapata. 2018. What's this movie about? a joint neural network architecture for movie content analysis. In *Proceedings of the NAACL-HLT*, pages 1770–1781.

Eun Young Ha, Joseph F. Grafsgaard, Christopher Mitchell, Kristy Elizabeth Boyer, and James C. Lester. 2012. Combining verbal and nonverbal features to overcome the "information gap" in task-oriented dialogue. In *Proceedings of the SIGDIAL*, pages 247–256.

Florian Hinterleitner, Georgina Neitzel, Sebastian Möller, and Christoph Norrenbrock. 2011. An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In *Proceedings of the Blizzard Challenge*.

HIT-TMG. 2022. dialogue-bart-large-chinese.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Yuxiang Jia, Huayi Dou, Shuai Cao, and Hongying Zan. 2021. Speaker identification and its application to social network construction for chinese novels. In *Proceedings of the IALP*, pages 13–18.

Mary Ritchie Key. 1977. Nonverbal communication: A research guide & bibliography.

Minchan Kim, Sung Jun Cheon, Byoung Jin Choi, Jong Jin Kim, and Nam Soo Kim. 2021. Expressive text-to-speech using style tag. In *Proceedings of the Interspeech*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the NMT*, pages 28–39.

Marianne LaFrance and Clara Mayo. 1978. Cultural aspects of nonverbal communication. *International Journal of Intercultural Relations*, 2(1):71–89.

Kathryn M Larsen and Charles Kent Smith. 1981. Assessment of nonverbal communication in the patient-physician interview. *J Fam Pract*, 12(3):481–488.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the NAACL-HLT*, pages 110–119.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the ACL*, pages 5755–5772.

Helmut Morsbach. 1973. Aspects of nonverbal communication in japan. *Journal of Nervous and Mental Disease*.

Fionn Murtagh, Adam Ganz, and Stewart McKie. 2009. The structure of narrative: the case of film scripts. *Pattern Recognition*, 42(2):302–312.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, CJ Linton, and Mihai Burzo. 2015. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the EMNLP*, pages 2336–2346.

Deepika Phutela. 2015. The importance of non-verbal communication. *IUP Journal of Soft Skills*, 9(4):43.

Fernando Poyatos. 1977. Forms and functions of nonverbal communication in the novel: A new perspective of the author-character-reader relationship. *Nonverbal Communication, Interaction, and Gesture*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the ACL*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the EMNLP*, pages 2383–2392.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the EMNLP*, pages 4035–4045.

Shane Saunderson and Goldie Nejat. 2019. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics*, 11:575–608.

Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. 2020. A simple and effective model for answering multi-span questions. In *Proceedings of the EMNLP*, pages 3074–3080.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint*, cs.CL/2109.05729v4.

Aron W Siegman. 1987. The telltale voice: Nonverbal messages of verbal. *Nonverbal behavior and communication*, page 351.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the NAACL-HLT*, pages 1768–1777.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the AAAI*, pages 4444–4451.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, and Claire Cardie. 2022. Improving machine reading comprehension with contextualized commonsense knowledge. In *Proceedings of the ACL*, pages 8736–8747.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging Chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

Jessica L Tracy, Daniel Randles, and Conor M Steckler. 2015. The nonverbal communication of emotions. *Current opinion in behavioral sciences*, 3:25–30.

Andrew Vassiliou. 2006. *Analysing film content: A text-based approach*. University of Surrey (United Kingdom).

Junshu Wang. 2017. *Information Extraction from TV Series Scripts for Uptake Prediction*. Ph.D. thesis, Auckland University of Technology.

Kanix Wang, Robert Stevens, Halima Alachram, Yu Li, Larisa Soldatova, Ross King, Sophia Ananiadou, Annika M Schoene, Maolin Li, Fenia Christopoulou,

et al. 2021. Nero: a biomedical named-entity (recognition) ontology with a large, annotated corpus reveals meaningful associations through text embedding. *NPJ systems biology and applications*, 7(1):38.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *Proceedings of the NLPCC*.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the CVPR*, pages 10687–10698.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the COLING*, pages 4762–4772.

Junjie Yang, Zhuosheng Zhang, and Hai Zhao. 2020. Multi-span style extraction for generative reading comprehension. *arXiv preprint*, cs.CL/2009.07382v2.

Dian Yu, Ben Zhou, and Dong Yu. 2022. End-to-end Chinese speaker identification. In *Proceedings of the NAACL-HLT*, pages 2274–2285.

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, and Yangqiu Song. 2022. ASER: towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. *Artificial Intelligence*, 309:103740.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. Aser: A large-scale eventuality knowledge graph. In *Proceedings of the WWW*, pages 201–211.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint*, cs.CL/1810.12885v1.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *Proceedings of the EMNLP-IJCNLP (System Demonstrations)*, pages 241–246.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI*, pages 730–738.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. *arXiv preprint*, cs.CL/2004.04100v1.

A Appendix

A.1 Clean NM Generation Data Construction

Following previous work (Vassiliou, 2006; Wang, 2017; Sun et al., 2022), we extract utterances and their corresponding speakers and NMs from scripts mostly using patterns.

A script is made up of multiple scenes. We first segment a script into scenes based on the blank lines and scene headings (e.g., EXT. (exterior spaces) and INT. (interior spaces)). In each scene, the utterance is set under its corresponding character, and the NM (if any exists) appears right after the character and is inside a parenthetical as shown in Table 10. We follow the aforementioned style format to design patterns to extract (utterance, speaker, NM) triples and the context such as previous utterances before the utterance.

Some scripts do not strictly follow the screenplay formats, and therefore an NM and its speaker may form a natural language sentence without any parentheticals as hints. We use a unified text-to-structure generation framework UIE (Lu et al., 2022), pretrained on large-scale structured and unstructured corpora. We manually annotate fourteen speaker identification instances (input: text that may contain speaker(s), NM(s), and an utterance; output: speaker of the utterance) and fine-tune UIE with them for few-shot learning. We discard instances with empty speakers and the ones with speakers of relatively confidence values (ranging from 0 to 1) less than 0.5 for quality control.[9]

To have a rough estimation of the quality of the extracted NMs from scripts, the authors manually check the randomly sampled 50 NMs in the dev set of Script, and the extraction accuracy (or exact match) is 100.0%. It is possible to carefully design more patterns or apply supervised methods to improve the recall of script-based NM extraction: for example, extracting those NMs written in the action lines in scripts that describe what the audience or readers are meant to see or hear in the scene, and they are not enclosed in parentheses, though this task itself requires additional annotation, which is beyond the scope of this paper.

A.2 Types of Nonverbal Messages

Kinesics: it contains body movement and facial expression (e.g., *"face grew grave"*), serving as an important form of nonverbal communication (Key, 1977).

Internal States: it contains two sub-types: intention (e.g., *"comfort"*, *"exhort"*, and *"patch up a lie"*) and inner feelings/emotions. The internal states are recognized to be associated with other types of nonverbal messages such as facial expressions (Tracy et al., 2015).

Pause: this refers to the pause that occurred before the target utterance was spoken (e.g., *"stunned for a while"* and *"without hesitation"*). Previous studies emphasize the role of pauses as they are considered as oral punctuation that conveys additional information (Cecot, 2001). This kind of information is crucial for speech applications such as audiobook reading (Hinterleitner et al., 2011).

Vocal-Related: based on analysis, we notice the majority of NMs are related to the characteristics of voice. We further categorize them into the following sub-types.

- **Addressee**: we only consider the cases where the addressee — the person at whom the speech is directed (Clark and Carlson, 1982) — is the speaker him/herself such as *"thought"* and *"talked to himself"*, we leave exploring of the impacts of the relationship (e.g., trusting) between the speaker and the addressee(s) on nonverbal messages (Larsen and Smith, 1981) to future work.

- **Number of speakers**: when the target utterance is spoken by multiple speakers. The NMs of each speaker may be constrained by others via social rules consciously or unconsciously (Gatica-Perez, 2009).

- **Tone**: we categorize an NM into this subtype (Ambady and Rosenthal, 1998) when there is no evidence in the texts to support that such an NM is observable through fa-

LYDIA *He's sure to be handsome.* **ELIZABETH** **(ironically)** *With five thousand a year, it would not matter if he had a big pink face.*
BINGLEY *Come Darcy, I must have you dance. I hate to see you standing by yourself in this stupid manner.* **MR DARCY** **(shakes his head)** *You know how I detest it.*

Table 10: Script sample with nonverbal messages from movie PRIDE AND PREJUDICE.

[9]We will release the extracted NM data based on scripts for research and non-commercial use.

cial expressions or is one of the inner feelings/emotions, for example, *"said coldly"* and *"said in amazement"*.

- **Speed**: the speed of the speech (e.g., fast and slow).
- **Volume**: it refers to the power of a speaker's voice, for example, *"in a soft voice"*.
- **Pitch**: in speech, it refers to the highness or lowness of a speaker's voice such as *"scream"* and *"with a deep voice"*. It is regarded as an important aspect of nonverbal communication, just as speed and volume (DeVito et al., 2000).
- **Timbre**: it refers to some important traits of the voice such as age, gender, and quality of the voice (e.g., *"hoarse voice"*), and these factors can also influence the nonverbal communication between speakers (Siegman, 1987).

Others: NMs that belong to none of the above categories such as dialect, singing, and languages.
NM Type Annotation: the authors first go through the data and design the above types and provide examples for each type in the guideline. When the annotators conduct the NM extraction annotation, to simplify the task, they will also select ONE type from the given thirteen candidate types for each selected NM. The inter-rater agreement measured by Cohen's kappa for NM typing is 0.73 based on the overlapped annotated spans by two annotators.

A.3 Human Annotation of NM Spans and Error Analysis of NM Extraction

As mentioned previously, we only keep those that occur **a short time before** the utterance is spoken or **at the same time** when there can exist several NMs in the context to ensure a high relevance between the target utterance and the annotated NMs. For example, given the preceding sentences:

> *Lu Wushuang and Cheng Ying immediately expressed their intentions of coming along with Huang Rong. They came out of Xiangyang, went around the enemy's camp, and went northwest. Huang Rong thought, "This time Xiang'er's intention is to find Yang Guo ..."*

verb phrases such as *"went around the enemy's camp"* and *"went northwest"* are not annotated as NMs of the speaker *"Huang Rong"* due to the unclear time interval based on the description.

Though NMs can be described in the context after the target utterance, we do not label those that are supported to occur after the target utterance. For example, *"flew forward"* is not regarded as an NM of the underlined target utterance based on the temporal clue *"after he said this"*, and *"said"* right after the speaker *"Zhang CuiShan"* of the target utterance is also ignored as it belongs to our defined relatively uninformative speech words.

> *Before Du DaJin could respond, sub-leader Shi cut in, "Just say what you want us to do." Zhang CuiShan said, "I'm going to break every single bone in your arms ...!" After he said this, he immediately flew forward.*

Error Analysis: to investigate the remaining challenges of NM extraction, we analyze the performance by the NM types (Table 11). There is still plenty of room for improvement to extract accurate NMs that are body movements and addressees.

We further manually check the best-performing generative extractor's predictions that are not exactly the same as the ground truth answers (EM=0). The common error types are (i) only one NM is extracted when multiple NMs exist (69.4%), (ii) the generated NM is incomplete compared with the single NM label (16.7%): for example, the pre-defined relatively uninformative word "道" (*"said"*) in "点点头道" (*"nodded and said"*) is not generated[10], and (iii) only the relatively uninformative NM is generated (5.6%). In particular, among the missed NMs, most of them belong to KINESICS (body movement 45.2% and facial expression 12.9%), followed by the feelings/emotions (25.8%) in INTERNAL STATES. For further improvement, it may be useful to increase the diversity of NME training data to involve books written by different authors or use in-context learning based on large language models such as GPT-3 (Brown et al., 2020) to generate weak-labeled multi-NM instances.

A.4 Data Statistics

A.5 Human Evaluation of NM Generation and Error Analysis

Human Evaluation: we define the following four metrics (M1–M4).

- **M1**: the RELEVANCE between the utterance and the NM based on the whole history context.
- **M2**: the FLUENCY of the NM. This metric mainly focuses on language expression such

[10] based on the annotation guideline introduced in Section 3.1

	EM↑	F1↑
body movement	74.5	90.7
facial expression	89.5	93.5
intention	78.9	90.7
feelings/emotions	84.6	93.8
pause	80.0	91.7
addressee	70.5	84.3
# speakers	100.0	100.0
tone	77.5	89.6
volume	78.9	90.2
speed*	100.0	100.0
pitch *	100.0	100.0
timbre*	100.0	100.0
others*	50.0	94.4

Table 11: The nonverbal message extraction performance by NM types on the dev set of the NME dataset ($\star$: fewer than 10 instances fall into this category).

	Script	Novel / Novel_L
type	human	pseudo-labeled*
supervision	direct	weak
# train instances	40K	397K / 749K
# train sources	404 scripts	521 novels
# dev	1,708	–
# dev sources	50 scripts	–
	CSK	$\text{LUGE}_\text{dialogue}$
type	human	human
supervision	indirect	indirect
# train instances	738K	13M
# train sources	–	four dialogue datasets
	C^3_D	EWECT
task	MRC	emotion classification
# train instances	5,856	27,768
# dev instances	1,825	2,000
# test instances	1,890	5,000
# sources	dialogues	microblogging

Table 12: Statistics of the nonverbal message generation data (pseudo-labeled: the NMs are extracted automatically by pattern-based, generative, or extractive NM extraction methods) and two dialogue/narrative understanding tasks. Novel is a subset of Novel_L.

as completeness, readability, fluency, free from grammar errors, etc.

- **M3**: the VALIDITY of the NM. A generated NM rated 1 in this metric can be associated with a type of NMs (e.g., fine-grained categories defined in Section A.2), and it can be factually sound considering the conversation scene description such as time and location in the given context.

- **M4**: the FACTUAL CONSISTENCY of the NM based on the context and the target utterance. In other words, if an NM has hallucination issues (Koehn and Knowles, 2017; Rohrbach et al., 2018) involving details (e.g., characters, scenes, and objects) not mentioned previously, it should be scored zero in this metric. For example, we can not infer the addressee's gender based on the given context in the last example in Table 7.

The scores for different metrics are required to be given **independently**. For example, given one utterance "好，你抓紧了，我让你冲上云霄。" (*"Okay, hold on tight, I'll let you soar to the heavens."*), the NM "看了看云霄" (*"look at the heaven"*) is less likely to be factually sound (M3=0) based on the given context (no descriptions about the fact that speakers are in or close to heaven) though it is rated 1 in M2 (fluency). Similarly, "拿起人参看了看说" (*"picked up ginseng and looked at it and said"*) itself is fluent, valid, and hallucination-free, a conflict exists between this NM and the given target utterance "哇，人参都不见了。" (*"Wow, all the ginsengs are gone!"*). The main difference between M1 and M3 is that we mainly focus on whether an NM can be conveyed by the target speaker in the current scene regardless of the utterances from other speakers in the same scene. For example, given the input "知福哥哥的学费总算凑齐了。[SEP]你们的恩情我永远忘不了。" (*"We finally made up the rest of the tuition for Zhifu."* [SEP] *"I will never forget your kindness."*), though *"sings"* is a valid NM (M3=1), it is irrelevant to the current context (M1=0).

Three commercial annotation teams participate in the evaluation. Given an NM generation instance (context, target utterance, NM_0, NM_1, NM_2, NM_3, NM_4) wherein NM comes from Script (i.e., ground truth) or is generated by models trained with different data, we randomly shuffle the NMs and hide the system label. For each NM in an instance, four metrics are rated (0 or 1) by three annotators from different annotation teams. The human agreement (κ) on the scores is measured using the average of any two annotation teams' Cohen's kappa (details in Table 13). For all four metrics, $\kappa = 0.55$ (moderate agreement). When we do not consider the hallucination issue in M4 as ground truth NM label cannot be judged using this metric, $\kappa = 0.64$ (substantial agreement). On average, each instance costs 0.39 RMB ($0.06).

	M1–M4	M1–M3
teams a and b	0.50	0.55
teams a and c	0.59	0.71
teams b and c	0.56	0.66
average	0.55	0.64

Table 13: Inter-rater agreement between different annotation teams for NM generation evaluation.

Error Analysis: based on the results of the best-

performing model (the last row in Table 5) on the dev set of Script, it may be helpful to further introduce commonsense knowledge from existing structured knowledge during training or via continual learning from large-scale books explicitly or implicitly. For example, the system generated NM *"points to a small bottle"* is less likely to happen compared with the ground truth *"picks up the cup"* (the first instance in Table 14), similar to the weighted edges in eventuality knowledge graph wherein the weight is defined as the frequencies of appearance of a piece of knowledge in the corpora (Zhang et al., 2020, 2022). In addition, including more books of diverse genres may alleviate some biases in script corpora. For example, NM *"to a walkie-talkie"* (see the full input in the second instance in Table 14) appears more frequently in the training set of Script than Novel_{L} (0.167% vs. 0.001%).

A.6 Evaluation of Verbal-Nonverbal Message Generation

We have shown that automatically extracted (context, utterance, NM) triples can benefit NM generation (Table 5). Another interesting question is whether the automatically extracted data can also be helpful when we aim to generate both an utterance and its corresponding NM, which is more challenging than NM generation. We use a similar formulation as that of the NM generation while we change the input to $U = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{k-1}\}$ and speaker s_k of utterance u_k, and we use the concatenation of u_k and $\mathbf{n} = \{n_1, n_2, \ldots, n_m\}$ as the output. We include the speaker information as there may exist multiple person entities in the previous context, and we are only interested in a certain speaker. We further process the labels by enclosing the nonverbal message with parentheses to generate structured results, motivated by the script formats (Figure 4). As a preliminary study, we experiment with T5 as the backbone model. We observe that, surprisingly, a model trained with weakly-labeled data constructed by any of our extractors can already achieve better performance than the same baseline trained with clean data (Table 15). However, we find that the lack of diversity is the main issue, reflected by the low DIST1/DIST2 scores. It might be useful to introduce more previous utterances or narratives as history context.

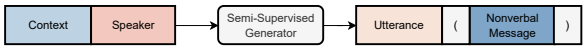

Figure 4: Verbal-nonverbal generation framework.

A.7 Enriched Examples of Downstream Tasks and Error Analysis

As introduced in Section 6.5, we add an NM right after each utterance or narrative in the original text input of a downstream task and keep all others the same. See examples for each task in Table 16.

We mainly analyze the wrong instances after the NMs are added to the original task inputs while they can be predicted correctly by the previous baseline. We notice that systems may be distracted by the newly added NMs especially when they are not highly relevant to the question and the conversation is short. For example, the NM *"smiles"* may distract the system from understanding the unspoken real intention (Table 17), though for human readers it seems that they do not intervene in the original expression.

A.8 Backbone Models, Hyper-Parameters, and Evaluation Metrics

We compare the backbone models in Table 19 and list hyper-parameters for three types of tasks in Table 18. We choose the two encoder models due to their superior performance in Chinese natural understanding tasks (Cui et al., 2021). We do not consider Chinese decoder models as publicly available models such as GPT-2[11] released by Zhao et al. (2019) are usually pre-trained on a relatively small scale corpus (Xu et al., 2020).

We follow extractive machine reading comprehension (span extraction) studies (e.g., (Rajpurkar et al., 2016)) using exact match and macro-averaged F1, which measures the average overlap between the extracted NM and the ground truth NM, both treated as bags of characters. We compute the average of the F1 over all of the NM extraction instances. A similar computation is conducted for the macro-averaged recall mentioned in Section 4.1. We use the NLTK tokenizer (Bird et al., 2009) if English words are also included in texts. We use the public evaluation code released by (Xu et al., 2020) to NM extraction.

Following previous dialogue response generation studies (e.g., (Celikyilmaz et al., 2020)), we use F1 as well as ROUGE-1 (measuring the overlap

[11] huggingface.co/uer/gpt2-chinese-cluecorpussmall.

input sequence (context [SEP] target utterance)	prediction	gold label
白雪梅进屋:你找我什么事[SEP] 我们这儿只有两个老师，曹老师有事回家去了。喝水。 Bai Xuemei entered the room: What do you want from me? [SEP] We only have two teachers here. Teacher Cao has gone home due to some personal reasons. Drink water.	指着一个小瓶 points to a small bottle	拿起杯子 picks up the cup
首领看到甲板前面的高墙上趴着五个人，激光就是从他们手里的枪里射出来的。[SEP]别开枪。你们是谁? The leader saw five people lying on the high wall in front of the deck, and the laser was shot from the guns in their hands. [SEP] Don't shoot. Who are you?	对着步话机 to a walkie-talkie	看向高处 looks up

Table 14: More NM generation system outputs.

train	model	type	F1	ROUGE-1	ROUGE-L	DIST-1	DIST-2
$\text{Script}_{\text{train}}$	T5_{base}	direct	7.09 (0.12)	22.96 (0.30)	20.81 (0.23)	3.43 (0.11)	12.70 (0.44)
$\text{Novel}_{\text{pattern}}$	T5_{base}	weak	7.10 (0.10)	23.71 (0.18)	20.81 (0.16)	3.58 (0.09)	14.54 (0.32)
$\text{Novel}_{\text{extractive}}$	T5_{base}	weak	7.25 (0.07)	23.77 (0.15)	20.87 (0.17)	5.04 (0.09)	22.23 (0.37)
$\text{Novel}_{\text{generative}}$	T5_{base}	weak	7.28 (0.06)	23.74 (0.14)	21.00 (0.11)	5.14 (0.12)	22.40 (0.45)

Table 15: The utterance-nonverbal generation performance on the dev set of the Script data (averaged score over five runs with different seeds).

C^3_{D}		
	Chinese Content	**English Translation**
document	男: 他准备这次考试很长时间了，可没想到，他居然没及格。(叹息) 女: 真的吗，太可惜了。不知道是什么原因啊。(叹息)	Man: He's been preparing for this exam for a long time, but he didn't expect to pass it. (sigh) Woman: Really? What a pity. I don't know why. (sigh)
question	女的是什么态度?	What is the attitude of the woman?
choices	高兴 生气 惊讶 惋惜⋆	happy angry surprised pity ⋆
document	女：我今天在街上遇到刘小如了。(笑着说) 男：真的吗?你们有十年没见面了吧? (惊喜的说) 女：是啊，但她还是像读大学时那样年轻漂亮。(点头) 男：周末请她到家里坐坐吧。(笑着说)	Woman: I met Liu Xiaoru on the street today. (said with a smile) Man: Really? You haven't seen each other for ten years, right? (surprised) Woman: Yeah, but she's still as young and beautiful as she was in college. (nod) Man: Please invite her to sit at home on weekends. (said with a smile)
question	女的跟刘小如可能是什么关系?	What is the relation between the woman and Liu Xiaoru?
choices	邻居 同学⋆ 同事	neighbor classmate ⋆ co-worker
EWECT		
	Chinese Content	**English Translation**
	【山羊也爱玩水，冲浪不输人！！】OMG，简直了～～逆天了～～～(惊叹)	【Goats also like playing in the water, and their surfing skills are comparable to humans'！！】OMG, it's just amazing! (exclaim)
label	惊奇	surprise

Table 16: Modified instances from downstream dialogue/narrative understanding tasks C^3_{D}(⋆: correct answer option) and EWECT.

C^3_{D}		
	Chinese Content	**English Translation**
document	女：您看这件衣服挺不错的，质量好，价钱也不贵。(笑着说) 男：再看看吧。(笑着说)	Woman: You see, this dress is very nice, of good quality and not expensive. (said with a smile) Man: Maybe I will check others. (said with a smile)
question	这个男的是什么意思?	What does this man mean?
choices	不想要这件⋆ 衣服挺好的 衣服太贵了 衣服质量不好	does not want this one ⋆ the cloth is great the cloth is expensive the cloth is of poor quality

Table 17: Negative example from the C^3_{D} dataset showing that adding NMs in the current way is not always helpful (⋆: correct answer option).

task	model	# of epochs	LR	batch	MaxLen	TSLen
NM extraction	RoBERTa-wwm-ext-large	5	3e-5	32	512	–
	$\text{MacBERT}_{\text{large}}$	5	3e-5	32	512	–
	T5_{base}	5	3e-4	64	512	20
	$\text{BART}_{\text{large}}$	5	2e-5	64	512	20
	$\text{DialBART}_{\text{large}}$	5	2e-5	64	512	20
NM generation	T5_{base}	1	3e-4	64	512	8
	$\text{BART}_{\text{large}}$	1	2e-5	64	512	8
	$\text{DialBART}_{\text{large}}$	1	2e-5	64	512	8
MRC	RoBERTa-wwm-ext-large	8	2e-5	64	512	–
emotion classification	RoBERTa-wwm-ext-large	5	3e-4	64	512	–

Table 18: Hyper-parameters settings for different fine-tuning tasks (LR: learning rate; TSLen: target sequence length).

of unigrams between the reference and hypothesis texts) and ROUGE-L (Lin, 2004) (measuring the longest matching sequence using longest common subsequence) to message the overlap between the generated NM and the ground truth. A public Python library is used for computing ROUGE scores.[12]. We implement DIST-1 and DIST-2 — the number of distinct unigrams and bigrams divided by the total number of generated characters — following (Li et al., 2016) to evaluate the diversity of the generated text.

model	size	task	type
RoBERTa-wwm-ext-large	324M	E	encoder
$\text{MacBERT}_{\text{large}}$	324M	E	encoder
T5_{base}	231M	E, G	encoder-decoder
$\text{BART}_{\text{large}}$	406M	E, G	encoder-decoder
$\text{DialBART}_{\text{large}}$	406M	E, G	encoder-decoder

Table 19: Descriptions about the Chinese backbone models (E/G: extraction/generation).

[12]https://github.com/pltrdy/rouge.