Journal of Artificial Intelligence Research 73 (2022) 511-533 Submitted 03/2021; published 02/2022

CASA: Conversational Aspect Sentiment Analysis for Dialogue Understanding

Linfeng Song LFSONG@TENCENT.COM
Tencent AI Lab, Bellevue, WA, USA 98004

Chunlei Xin CLXIN@STU.XMU.EDU.CN
Shaopeng Lai SPLAI@STU.XMU.EDU.CN
Ante Wang WANGANTE@STU.XMU.EDU.CN
Xiamen University, Xiamen, Fujian, China 361005

Jinsong Su* JSSU@XMU.EDU.CN
Xiamen University, Xiamen, Fujian, China 361005
Pengcheng Lab, Shenzhen, Guangdong, China 518066

Kun Xu KXKUNXU@TENCENT.COM
Tencent AI Lab, Bellevue, WA, USA 98004

Abstract

Dialogue understanding has always been a bottleneck for many conversational tasks, such as dialogue response generation and conversational question answering. To expedite the progress in this area, we introduce the task of *conversational aspect sentiment analysis (CASA)* that can provide useful fine-grained sentiment information for dialogue understanding and planning. Overall, this task extends the standard aspect-based sentiment analysis to the conversational scenario with several major adaptations. To aid the training and evaluation of data-driven methods, we annotate 3,000 chit-chat dialogues (27,198 sentences) with fine-grained sentiment information, including all sentiment expressions, their polarities and the corresponding target mentions. We also annotate an out-of-domain test set of 200 dialogues for robustness evaluation. Besides, we develop multiple baselines based on either pretrained BERT or self-attention for preliminary study. Experimental results show that our BERT-based model has strong performances for both in-domain and out-of-domain datasets, and thorough analysis indicates several potential directions for further improvements.

1. Introduction

Modeling chit-chat dialogues has been attracting increasing research attention due to its potential values in facilitating human-computer communications. Until now, most studies (Vinyals and Le, 2015; Sordoni et al., 2015; Li et al., 2016; Serban et al., 2016, 2017; Zhao et al., 2017; Shao et al., 2017; Xing et al., 2018) on building dialogue systems have mainly focused on designing end-to-end neural models that only consume the features from surface strings. These systems have shown to be successful for producing single-turn responses. However, they are still far from satisfactory for multi-turn scenarios. Specifically, these models suffer from several severe problems, such as (1) they behave passively during conversations, producing safe yet trivial responses (e.g. "*I don't know*" and "*It's Okay*") (Jiang

* Corresponding author

A:	Did you see the match between Argentina and Brazil last night ?
B:	Yes , of course .
A:	Messi contributed two goals , while Neymar got nothing .
B:	Yeah, the goalkeeper was just stunned .
A:	His overall performance is terrific !
B:	Yes , it is .

Table 1: A three-turn dialogue, where a sentiment expression (e.g. "*terrific*") and its mention (e.g. "*Messi*") are labeled with the same type of underline.

and de Rijke, 2018; Wu et al., 2019), and (2) their generated multi-turn responses are often barely coherent with each other (Mei et al., 2017).

To alleviate these problems, later research explored *external* knowledge as additional background information to aid dialogue understanding. For instance, the effects of knowledge graphs (KG) (Kumar et al., 2019; Wu et al., 2019), commonsense knowledge (Young et al., 2018), personalities (Zhang et al., 2018; Lian et al., 2019) and emotions (Colombo et al., 2019; Zhong et al., 2019; Asghar et al., 2018) have been investigated to produce engaged and meaningful responses. However, these types of external knowledge, such as a related KG or a proper emotion for the next response, are usually not explicitly expressed in dialogues, and thus human annotations are required along with the benchmark datasets. As a result, this may lead to a mismatch problem, as the external information may be unavailable or difficult to obtain for real-world applications. For example, it is often difficult to detect a related KG when discussing the personal situations of the user. Besides, leveraging KGs in practical scenarios requires entity linking as a prerequisite step, which often introduces additional errors.

In this paper, we explore another direction that extracts the *internal* knowledge from each dialogue, which can be exploited to benefit dialogue understanding. To this end, we propose a new task: *conversational aspect sentiment analysis (CASA)*, which aims to extract user opinions (and polarity) and the corresponding mentions from dialogues. Our intuition is that humans often express their emotions on the entities (e.g. a recent soccer match or a famous NBA player) that they are talking about during conversations, and thus CASA can provide helpful features for general-domain dialogue understanding. More specifically, accurately extracting people's opinions and the corresponding entities from dialogue histories can help chatbots plan subsequent topics, making them more engaging and active for multi-turn conversations. For instance, if a user mentions that he/she is a super fan of the soccer player "*Lionel Messi*", then a chatbot can mention the recent news of Messi. One may argue that aspect sentiments could also be implicitly captured by the vectorized representations from a dialogue encoder. However, explicitly extracting this information can help to alleviate data sparsity, as the same aspect term and sentiment pair can appear in various contexts. Besides, explicitly representing this information can increase model interpretability, making it easier to be combined with other knowledge (e.g. an external KG). Taking the same example of "*Lionel Messi*", by combining the parsing results of CASA with external KGs, the chatbot may even recommend the recent match of Messi's soccer club "*Futbol Club Barcelona*".

While there have been extensive studies on aspect sentiment analysis, accurately identifying fine-grained sentiments in dialogue conversations still remains challenging. One reason is that most of the datasets (Dong et al., 2014; Pontiki et al., 2016; Saeidi et al., 2016; Jiang et al., 2019) for standard aspect sentiment analysis contain very limited numbers (typically thousands) of instances, and they only cover a few domains (such as reviews of hotels and restaurants), while daily conversations are open-domain. Besides, in these datasets, a sentiment expression is usually close to its corresponding aspect terms within a short sentence. Conversely, the distances can be beyond several dialogue turns for conversations, where frequent ellipsis and anaphora introduce more complexity for reasoning. For example in Table 1, the mention "*Messi*" appears in the third utterance, while the corresponding sentiment word "*terrific*" is in the fifth utterance. Moreover, "*Neymar*" introduces further challenges as a highly confusing candidate mention. This is just a 3-turn instance, let alone the increasing complexity with more turns. In order to boost the research for CASA and dialogue understanding, we manually annotate a popular Chinese dataset of chit-chat dialogues (Wu et al., 2019) with mentions and sentiment expressions for training and evaluating models. Chinese is a flexible language, where both ellipsis and anaphora occur frequently. Therefore, it can be a good test-bed for conducting research on conversational sentiment analysis. Our dataset contains 3,000 dialogues, including roughly 27,200 sentences, 8,000 mentions and their sentiment expressions. We also create an additional test set containing 200 dialogues in other domains to verify model robustness.

To efficiently model CASA, we decompose the task into two subtasks: (1) *sentiment extraction (SE)* and (2) *mention extraction (ME)*. Specifically, SE aims to find all sentiment expressions from the last user utterance and determine the polarity of each extracted expression. Meanwhile, ME is to extract the corresponding mention from the dialogue history for each sentiment expression. We cast SE as a sequence labeling task under the BIO scheme and treat ME as a span prediction problem (such as reading comprehension, Rajpurkar et al. 2016), and we adopt models using BERT (Devlin et al., 2019) or regular self-attention (Vaswani et al., 2017) to solve each of them. Similar architectures (Sun et al., 2019; Yang et al., 2019) have shown state-of-the-art performances for standard aspect sentiment analysis. As ME requires understanding the whole dialogue history, we also leverage other types of rich features, such as the information of sentence-wise distances and speaker IDs, to aid the modeling of long-distance dependencies.

Experiments show that our BERT model achieves decent performances on both the DuConv (in-domain) test set and the out-of-domain test set. In-depth analysis further demonstrates that the errors can be mainly attributed to several specific categories, and thus there still remains large room for further improvements.

We further evaluate the usefulness of CASA on knowledge-driven dialogue response generation, where each dialogue discuses a given document. Taking Transformer-based baselines with or without full document, we can observe that CASA can help improve both memory efficiency and response quality by selecting only relevant knowledge from provided news documents.

Overall, our contributions in this paper can be summarised as follows:

- We introduce the task of *conversational aspect sentiment analysis* as a further step for dialogue understanding. It can provide fine-grained information about the opinions of users towards the entities mentioned in their dialogues.

- We manually create a dataset of 3,000 dialogues (with 27,198 sentences) for training and evaluating data-driven methods. This can boost the research for sentiment analysis as well.
- We introduce strong baseline models for a preliminary study of this task and release our code and datasets at `https://github.com/freesunshine0316/lab-conv-asa`.
- On a knowledge-driven dialogue response generation task, we demonstrate that CASA can accurately select relevant knowledge, improving both effectiveness and memory usage.

2. Related Work

The related work mainly consists of dialogue understanding and aspect sentiment analysis, which will be described as follows:

2.1 Dialogue Understanding

How to effectively represent (understand) dialogue histories has always been an open question. Most recent efforts enhance dialogue understanding by exploring *external* knowledge, such as knowledge graphs (Kumar et al., 2019; Wu et al., 2019), commonsense knowledge (Young et al., 2018), and other resources (Colombo et al., 2019; Zhong et al., 2019; Asghar et al., 2018). On the other hand, little work has explored the types of *internal* features from dialogue itself. Notwithstanding, Xing et al. (2017) treat each dialogue utterance as a "document" and employ a latent topic model (Blei et al., 2003) to highlight the key words within dialogue histories for better response generation. Both our extracted mentions and the key words from Xing et al. (2017) can capture the important content information. However, our mentions only focus on the core entities in dialogues, having higher quality than the topic words from Xing et al. (2017) that can be any tokens. Besides, we make one step further by extracting user opinions and their polarities towards these entity mentions. Recently, people propose the task of conversational semantic role labeling (Xu et al., 2021) and dialogue-based relation extraction (Yu et al., 2020; Nan et al., 2021) to extract the semantic and relational structures across a dialogue. This is also one type of internal knowledge, which is intuitively orthogonal to the information extracted from our aspect-based sentiment analysis task.

2.2 Aspect-Level Sentiment Analysis

The existing studies (Saeidi et al., 2016; Pontiki et al., 2016; Jiang et al., 2019) on standard aspect sentiment analysis (SASA) mainly focus on the sentiments of product aspects, where each review is a short sentence and its sentiment expressions are usually close to their aspect terms. Also these datasets only provide the annotations of aspects and their corresponding polarities without annotating the sentiment expressions that indicate the polarities. There are some early datasets (Liao et al., 2013; Zhao et al., 2014; Kessler and Nicolov, 2009) with annotated sentiment expressions, yet they have far fewer instances than the recent ones. For the first time, we introduce CASA by extending SASA to the conversation scenario, where a sentiment expression can be in a different utterance from the corresponding mention

(a)	A:	[**a** 一些(some) 经典的(classic) 电影(movie)] 就是(are) 那么(that) [**a** 百看不厌(not boring even after watching 100 times) **+1**] 。
	B:	你(you) 指(mean) 哪些(which ones) ?
	A:	比如(for example) [**a** 阿甘正传(Forrest Gump)] 。
(a)	A:	我(I) 昨晚(last night) 又(again) 看了(watched) [**a** 肖申克的救赎(The Shawshank Redemption)] ， 简直(so) [**a** 百看不厌(not boring even after watching 100 times) **+1**] 。
	B:	是的(yes) ， 它(it) 是(is) [**a** 几代人(several generations) 的('s) 经典(classic) **+1**] 。
	A:	不过(but) ， 这部(this) 剧(movie) 的('s) 主演(protagonists) ， 我(I) 觉得(think) [**b** 蒂姆 · 罗宾斯(Tim Robbins)] 的('s) 演技(acting skill) [**b** 不(not) 算(is) 那么(that) 出众(outstanding) **-1**] 。
(c)	A:	昨天(yesterday) 看了(watched) [**a** 冰雪奇缘2(Frozen II)] ， 还是(is) [**a** 非常(very) 不错的(good) **+1**] ， 虽然(although) 没有(not) 超越(surpass) [**b** 第一部(the first version)] 。
	B:	第一部(the first version) 已经(already) 是(is) [**b** 经典(classic) **+1**] ， 不(not) 那么(that) 容易(easy) 被(be) 超越(outperformed) 。
	A:	是啊(yes) ， 要不(that's) 怎么(why) 力压(outperform) 神偷奶爸2(Despicable Me 2) 拿了(got) 当年的(that year) 奥斯卡奖(Oscar award) 。
	B:	那首(that) let it go 堪比(comparable) 我心永恒(my heart will go on) 。

Table 2: Some annotated dialogue sessions from DuConv, where we provide an English translation for each Chinese word for better understanding. Sentiments and mentions are labeled with "[ ]" brackets.

(a.k.a. aspect). Besides, we provide a new large-scale dataset with the annotations of mentions, their polarities and the supporting sentiment expressions to support the training and evaluation of state-of-the-art data-driven methods.

In addition, the existing work on SATA (Nguyen and Shirai, 2015; Ruder et al., 2016; Wang et al., 2016; Chen et al., 2017; Wang et al., 2017; Li et al., 2019; Luo et al., 2019) typically follows the assumption that product reviews are usually aspect centric, and that each aspect corresponds to one and only one sentiment expression. Conversely, we remove this assumption. It is because people can give different sentiment expressions towards one mention, and people can talk about many things, while they only express subjective opinions on a few of them.

3. Data Annotation

We annotate the mentions and their sentiment expressions on *DuConv* (Wu et al., 2019), a popular public dataset of Chinese chit-chat dialogues, and *NewsDialogue* (Wang et al., 2021), an dataset of dialogues discussing news. These two datasets cover different domains. We choose DuConv as the major dataset to label most of our instances, while the annotated examples from NewsDialogue just serve as additional test instances to evaluate the model robustness on domain adaptation.

3.1 Dialogue Datasets

The dialogues of DuConv mainly discuss movies, TV shows and celebrities, containing a large number of opinions. It consists of around 30K dialogues with 270K dialog turns, where each dialogue is accompanied with a knowledge graph fragment containing the entities and other relevant information mentioned in the dialogue. For each dialog session, two crowd-sourced annotators are hired to conduct a multi-turn conversation discussing the entities and relations within a given KG, where one person plays the role of the conversation leader and the other acts as the follower.

On the other hand, the NewsDialogue dataset contains 20K dialogues with 400K dialogue turns and covers broader domains, including sports, technology, education and so on. Each dialogue from this dataset is generated by requiring two crowd workers to discuss a piece of recent news (e.g. about a basketball game or a new released 5G chip) and its related topics. In this work, this dataset can serve as a test-bed for evaluating model robustness.

3.2 Annotation Details

We randomly select 3,000 and 200 dialogues from DuConv and NewsDialogue for annotating sentiment expressions and corresponding mentions. As for the details of the annotation procedure, we first ask two annotators to label each dialogue independently, and then they resolve any annotation disagreements by themselves according to our annotation guidelines shown below. Table 2 illustrates several cases that reflect our annotation criterion. Generally, each target mention is surrounded by "[$\boldsymbol{x}$]", where $\boldsymbol{x}$ is a variable to distinguish different mentions, and possible values can be any lowercased English characters. The corresponding sentiment expressions for mention $\boldsymbol{x}$ are surrounded by "[$\boldsymbol{x}$ $\boldsymbol{s}$]", where $\boldsymbol{s} \in \{-1, 0, +1\}$ represents the sentiment value. For any sentiment expression that refers to multiple mentions (e.g. $\boldsymbol{x}$ and $\boldsymbol{y}$), all corresponding mentions are considered ("[$\boldsymbol{x}$+$\boldsymbol{y}$ $\boldsymbol{s}$]").

Our annotation procedure mainly follows the guideline[1] of the SemEval-2014 benchmark (Pontiki et al., 2014) for standard aspect-based sentiment analysis. In addition, we make a few adaptations for our conversational scenario. One major change is that we annotate not only the polarity values but also the corresponding sentiment expressions for each mention. Taking Table 1 as an example, we first annotate "*Messi*" and "*Neymar*" as mentions. Then we label "*terrific*" as a positive expression to "*Messi*". Please note that it is necessary to annotate all sentiment expressions because people can have different opinions towards the same entity mention, and their opinions can even change during a conversation.

Another major adaptation is that we directly consider each mention (e.g. "*Messi*" in Table 1) as the target of sentiment expressions, rather than the detailed aspect term or action of the mention (e.g. "*the long shot*"). This is because mentions are usually the "central topics" of conversations, and a finer granularity may cause the data sparsity problem when training data-driven models. Besides, we skip annotating the mention categories, as it is very difficult to categorize the mentions in our conversations from various domains. By contrast, the SemEval benchmark provides such annotations because it only focuses on one or two domains (e.g. restaurant), where the possible categories (e.g. service, food and

1 `http://alt.qcri.org/semeval2014/task4/data/uploads/semeval14_absa_annotationguidelines.pdf`

price) are easily enumerable. In addition, we stick to the following standards for annotating mentions:

- Each annotated mention should be as specific as possible. For instance, in dialogue (a) of Table 2, we choose "一些(some) 经典的(classic) 电影(movie)" as the mention, rather than "电影(movie)" or "经典的(classic) 电影(movie)".
- For multiple mentions that correspond to the same entity, we only annotate the most specific ones. Taking dialogue (b) as an example, we only select "肖申克的救赎(The Shawshank Redemption)", ignoring other coreference pronouns, such as "它(it)" and "这部(this) 剧(movie)".
- If there is any corresponding sentiment expression (e.g. "百看不厌(not boring even after watching 100 times)" in dialogue (a)) in an earlier turn than the already annotated mentions ("阿甘正传(Forest Gump)" in dialogue (a)), we also annotate the very first mention ("一些(some) 经典的(classic) 电影(movie)") of the entity, no matter whether the mention is specific enough. This can reduce the situation where no target mention is available for a newly occurred sentiment expression when processing a dialogue session.
- The mentions without the corresponding sentiment expressions (e.g. "神偷奶爸2 (Despicable Me 2)" in dialogue (c) of Table 2) are not annotated. This is because it may be inaccurate to infer the sentiments for these mentions without explicit sentiment expressions.

We also adhere to the following rules when annotating sentiment expressions:

- We define sentiment expressions as subjective opinions, so expressions describing facts (e.g. "拿了(got) 当年的(that year) 奥斯卡奖(Oscar award)" in dialogue (c) of Table 2) are not considered as sentiment expressions.
- Our sentiment expressions do not include the situations requiring background knowledge or reasoning. Taking the last sentence in dialogue (c) of Table 2 as an example, we do not annotate "堪比(comparable) 我心永恒(my heart will go on)" as a sentiment expression, because this requires the background knowledge that "我心永恒(my heart will go on)" is a very famous song.
- Similar to mentions, each annotated sentiment expression is required to be specific enough. As shown in dialogue (b) of Table 2, we annotate "几代人(several generations) 的('s) 经典(classic)" rather than "经典(classic)", because "几代人(several generations)" indicates the extent of the sentiment. But, we do not include "简直(so)" into "百看不厌(not boring even after watching 100 times)", as it does not significantly contribute to make the sentiment more specific.

3.3 Data Statistics

Table 3 provides some statistics for our annotated datasets, where "cross turn %" represents the percent of sentiment expressions whose target mentions are not in the same dialogue

Dataset	#Dialog (#Sent)	#Sentiment (cross-turn %) Pos % Neu % Neg %	#Mention (Avg. len)
DuConv	3,000 (27,198)	10,436 (46.50%) 81.74% 4.83% 13.43%	8,009 (2.25)
NewsDialogue	200 (4,008)	1,183 (66.02%) 73.88% 3.80% 22.32%	587 (1.92)

Table 3: Statistics for annotated datasets. #X represents "the number of X", and x% ($x \in$ {Pos, Neu, Neg}) shows the percent of sentiment expression belonging to each polarity value.

Dataset	Frequent Mentions	Frequent Sentiment Expressions
DuConv	电影 (movie, 24.1%); 一部电影 (one movie, 11.7%); 这部电影的导演 (the director of this movie, 5.8%); 明星 (star, 5.3%); 导演 (director, 5.1%); 喜剧 (comedy, 3.6%); 剧情电影 (drama movie, 3.6%); 纪录片 (documentary, 3.5%); 主演 (starring, 3.2%); 人间喜剧 (The Human Comedy, 3.0%); 阿凡达 (Avatar, 2.7%); 爱情公寓 (iPartment, 2.6%); 美国电影 (American movie, 2.5%);	喜欢 (like, 18.2%); 很喜欢 (like very much, 10.0%); 很不错 (very good, 9.0%); 很好看 (very nice, 7.2%) 口碑很差 (very bad rating, 6.3%) 还不错 (good, 6.0%); 口碑不错 (good rating, 4.5%) 很有才华 (very talented, 2.7%) 还可以 (okay, 2.1%) 不是很了解 (not familiar, 2.1%)
NewsDialogue	足球 (football); 女排 (women's volleyball, 6.9%); 梅西 (Messi, 5.6%); 贝弗利 (Beverly, 4.2%); 苹果 (Apple Inc, 4.2%); 篮球 (basketball, 4.2%); 恒大 (Evergrande, 4.2%); 快船 (Clippers, 4.2%); 巴萨 (FC Barcelona, 4.2%); 厄文 (Kyrie Irving, 4.2%); 华为 (Huawei, 4.2%); 中国足球 (Chinese Football); NBA (NBA, 4.2%); 阿森纳(Arsenal F.C., 2.8%); 阿尔萨德(Al Sadd SC, 2.8%)	喜欢(like, 16.5%); 不错(good, 11.2%); 很喜欢(like very much, 8.5%); 很厉害(awesome, 6.4%); 挺好(very good, 5.9%); 挺厉害(awesome, 5.9%); 很好看(looks good, 3.7%); 太帅(so handsome, 2.7%); 优秀(excellent, 2.7%); 不太了解(not familiar, 2.7%) 还不错(okay, 2.1%)

Table 4: Frequent mentions and sentiment expressions in both datasets, where the English translations and frequencies are in parentheses.

turn. Out of the 27K sentences in DuConv, we find more than 10K sentiment expressions, and nearly half (46.50%) of them describe mentions in different turns. As for the sentiment expressions, more than 80% are positive, and only less than 5% are neutral. Intuitively, neutral expressions are less informative than the other two, thus they are rarely used by humans.

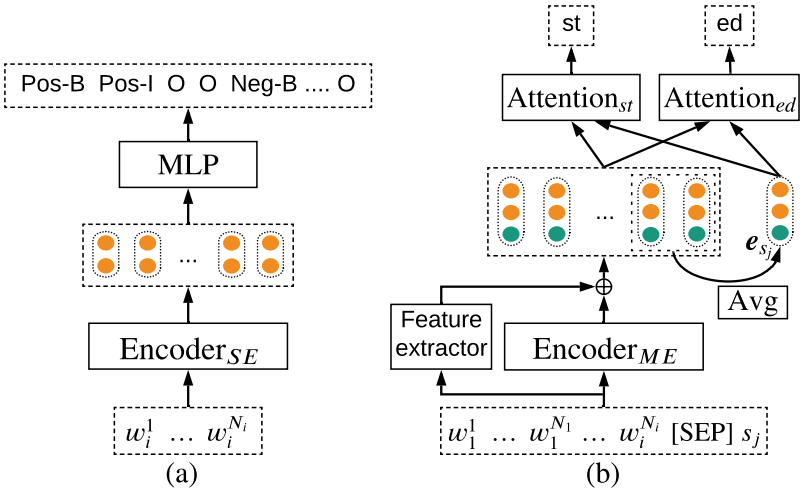

Figure 1: Model architectures for (a) sentiment extraction, and (b) mention extraction. Modules (e.g., *"Encoder$_{SE}$"*) are surrounded by solid lines, while data and outputs are boxed with dashed lines. The s_j represents a sentiment expression as defined in §4.1.

Comparatively, NewsDialogue has a higher percent of cross-turn sentiment expressions than DuConv (66.02% vs 46.50%), which makes this dataset more challenging than DuConv. Besides, it has a higher percent of negative sentiment expressions (22.32% vs 13.43%), and on average there are more sentiment expressions that discuss the same mention. On both datasets there are fewer mentions than sentiment expressions (8,009 vs 10,436, and 587 vs 4,008 for DuConv and NewsDialog, respectively), because humans can make multiple comments for one entity.

In the subsequent experiments, we split DuConv into training, development and test sets, with each containing 80%, 10% and 10% of the whole data, respectively. The whole NewsDialogue dataset is used as an out-of-domain test set. Regarding sentiment expressions, the DuConv testset and NewsDialog have 27.13% and 77.09% unseen expressions, respectively, and the numbers are 61.27% and 95.74% for mentions. The high percent of unseen cases of NewsDialog makes it a suitable test set for out-of-domain evaluation.

Table 4 lists the top frequent mentions and sentiment expressions of both datasets. For mentions, as previously mentioned, both datasets cover very different domains, where DuConv mainly discusses movies while NewsDialogue mostly focuses on sports (e.g., football, basketball and volleyball). This creates a good testbed for out-of-domain evaluation on mention detection. On the other hand, sentiment expressions in both datasets are similar. The reason can be the nature of human language, where the sentiment words are usually domain invariant. Besides, we observe that most sentiment words are positive, which is consistent with the statistics shown in Table 3.

4. Model

In this section, we first formally introduce the task of CASA, before giving descriptions on several models for this task. As shown in Figure 1, we design strong models based on

self-attention (Vaswani et al., 2017) and pretrained BERT (Devlin et al., 2019) to solve each of the sub-tasks. Recent work (Sun et al., 2019; Yang et al., 2019) has shown that similar model architectures can demonstrate highly competitive performances for sentiment tasks.

4.1 Task Definition

Formally, the input of our CASA task can be represented as a list of dialogue utterances $X_1, X_2, \ldots, X_i$, where $X_i = w_i^1, \ldots, w_i^{N_i}$ is the latest dialogue turn and N_i represents the length of X_i. The goal of CASA can be separated into two sub-tasks: (1) extracting all sentiment expressions $\{s_1, \ldots, s_M\}$ and their polarity values $\{p_1, \ldots, p_M\}$ (*sentiment extraction, SE*) from X_i, and (2) detecting their corresponding mentions $\{m_1, \ldots, m_M\}$ (*mention extraction, ME*) from the whole dialogue history $X_1, X_2, \ldots, X_i$. Each sentiment expression s_j can be a word or a phrase within turn X_i, and its polarity value p_j is chosen from three possible values: -1 (negative), 0 (neural) and +1 (positive).

4.2 Model for Sentiment Extraction

As shown in Figure 1(a), we design a model (Encoder$_{SE}$) that treats extracting sentiment expressions and detecting their polarities together as sequence labeling. Using BIO scheme, we consider the tag set that contains seven possible tags.[2] The model adopts either a pretrained BERT (Devlin et al., 2019) or multiple self-attention (Vaswani et al., 2017) layers to generate context-sensitive embeddings for an input sentence $w_i^1, \ldots, w_i^{N_i}$:

$$\boldsymbol{b}_i^1, \ldots, \boldsymbol{b}_i^{N_i} = \texttt{Encoder}_{SE}(w_i^1, \ldots, w_i^{N_i}) \tag{1}$$

For both types of encoders, we associate vectors to sub-word units based on the byte pair encoding (BPE) (Sennrich et al., 2016) to avoid out-of-vocabulary (OOV) tokens. We then follow Kitaev et al. (2019) to obtain word-level representations from sub-word units by averaging the vectors of the sub-word units within each word. The word representations are then fed into a multi-layer perceptron (MLP) with 7 output units and softmax activation to predict the tag for each input word (e.g. w_i^k). Next, all sentiment expressions and their polarities are inferred from the tags. For example, the model is supposed to produce tags "*O O O O Pos-B O*" for the sentence "*His overall performance is terrific !*" in Table 1, and a positive sentiment expression "*terrific*" is then detected from the tags.

4.3 Model for Mention Extraction

Given the results of sentiment extraction, a mention extractor (Encoder$_{ME}$) in Figure 1(b) is adopted to extract the corresponding mention m_j for each sentiment expression s_j. It takes the concatenation of all dialogue turns $w_1^1, \ldots, w_i^{N_i}$ and the associated expression s_j as inputs, leveraging another encoder based on self-attention or pretrained BERT to obtain their contextual embeddings:

$$\boldsymbol{b}_1^1, \ldots, \boldsymbol{b}_i^{N_i}, \boldsymbol{b}_{\texttt{SEP}}, \boldsymbol{b}_{s_j}^1, \ldots, \boldsymbol{b}_{s_j}^{|s_j|} = \texttt{Encoder}_{ME}(w_1^1, \ldots, w_i^{N_i}, [\texttt{SEP}], w_{s_j}^1, \ldots, w_{s_j}^{|s_j|}), \tag{2}$$

where $w_{s_j}^1, \ldots, w_{s_j}^{|s_j|}$ represent the tokens of the sentiment expression s_j, and [SEP] is an artificial token to separate a context and a sentiment expression. Similar with sentiment

2 {Pos, Neu, Neg} × {B,I}, plus O

extraction, we also use the vectors of last sub-word units to obtain word-level representations. Compared with the sentiment extraction task (Section 4.2), this task requires longer-distance reasoning throughout the whole dialogue. To address this issue, we introduce rich features including turn-wise distances and speaker information to model the cross-sentence correlations. Specifically, the turn-wise distances are relative distances to the current turn that are bucketed into $[0, 1, 2, 3, 4, 5+, 8+, 10+]$. The speaker information is a binary feature indicating whether a token in the dialogue history is from the same speaker as the current turn. Both types of information are represented by embeddings (e.g. $\boldsymbol{d}_k^l$ and $\boldsymbol{s}_k^l$ correspond to the distance embedding and the speaker embedding for token w_k^l), before being concatenated with the encoder outputs (i.e. $\boldsymbol{b}_k^l$) to obtain the rich contextual representations: $\boldsymbol{e}_k^l = \boldsymbol{b}_k^l \oplus \boldsymbol{d}_k^l \oplus \boldsymbol{s}_k^l$, where $\oplus$ denotes the concatenation operation.

Next, a vector representing the whole sentiment expression s_j is generated by averaging the contextual representations of all tokens within it: $\boldsymbol{e}_{s_j} = \texttt{Avg}(\boldsymbol{e}_{s_j}^1, \ldots, \boldsymbol{e}_{s_j}^L)$, where L is the length of s_j. Taking the vector ($\boldsymbol{e}_{s_j}$) for the sentiment expression and the concatenated dialogue-history representation ($\boldsymbol{E}_{diag} = [\boldsymbol{e}_1^1; \ldots; \boldsymbol{e}_i^{N_i}]$) as the query and memory, we then adopt two attention layers (Bahdanau et al., 2015) to calculate the distributions for the start and end boundaries of the target mention, respectively. The overall distribution for the target mention is defined as the product of both distributions:

$$\boldsymbol{\phi}_{st} = \texttt{Attention}_{st}(\boldsymbol{E}_{diag}, \boldsymbol{e}_{s_j}), \tag{3}$$

$$\boldsymbol{\phi}_{ed} = \texttt{Attention}_{ed}(\boldsymbol{E}_{diag}, \boldsymbol{e}_{s_j}), \tag{4}$$

$$\boldsymbol{\phi} = \boldsymbol{\phi}_{st}^{\intercal} \boldsymbol{\phi}_{ed}. \tag{5}$$

Finally, the target mention (st, ed) is produced by choosing both boundaries st and ed that yield the highest score from $\boldsymbol{\phi}[st, ed]$, where $st \leq ed$ and they need to be in the same utterance.

4.4 Training

Both models are trained with the standard negative log-likelihood loss over gold references y: $-\log p(y|X)$, where X represents the task input, and y corresponds to the reference. For sentiment extraction, y is tags in the BIO scheme, while for mention extraction, y corresponds to the boundaries of each gold span.

5. Experiments

We conduct experiments on our annotated datasets (more details are shown in §3.3 and Table 3) to investigate the current performances on dialogue-based sentiment analysis and to pinpoint the main challenges for future improvements.

5.1 Settings

Following most existing work on standard aspect sentiment analysis, we report F1 scores of extracting sentiment expressions. For extracting the corresponding mentions given sentiments, we report accuracy scores of fully correct mentions, where each mention is represented as a span in the dialogue history. For BERT-based models, we adopt the hug-

Encoder	DuConv (in-domain)				NewsDialogue (out-of-domain)			
	SE	$SE_{unlabeled}$	ME	ME_{cross}	SE	$SE_{unlabeled}$	ME	ME_{cross}
Self-Attention	55.67	55.97	26.14	20.39	20.72	21.13	22.32	18.71
BERT-freeze	71.95	72.20	50.55	58.39	43.90	44.42	39.31	40.03
BERT	78.44	78.89	79.08	80.95	52.39	52.78	63.97	58.65

Table 5: Main results, where "SE" and "ME" represent the performances on sentiment extraction (F1 score) and mention extraction (accuracy) tasks, respectively. "$SE_{unlabeled}$" indicates the F1 scores that do not consider the polarities of extracted expressions, and "ME_{cross}" shows the accuracies for only the cross-turn situations.

gingface[3]-based `RoBERTa-wwm-ext` model (Cui et al., 2020). For the models based on self-attention, four layers are stacked, and we choose randomly initialized embeddings with 768 units for sub-word units to be consistent with our BERT model. All models are trained for 20 epochs using Adam (Kingma and Ba, 2014) with learning rate 10^{-5}.

5.2 Results

Table 5 shows the main results of our models, where *BERT* gives the best scores. When freezing the BERT parameters (*BERT-freeze*), we observe significant performance drops, indicating the necessity of finetuning BERT. One major reason can be the domain shift problem, as conversations usually have very different genre from the narrative sentences for large-scale pretraining. Besides, our data contains very recent entities that are excluded from the narrative sentences for large-scale pretraining. There is another large performance drop by changing parameter-freeze BERT to multiple self-attention layers (*Self-Attention*) with randomly initialized embeddings as the encoder, which indicates the usefulness of large-scale pretraining. These performance changes are also consistent on the out-of-domain NewsDialogue test set as well. Overall, using and finetuning BERT are important factors for reaching a descent performance.

Regarding our BERT model, on the in-domain DuConv test set, it achieves reasonable performances of 78.44 and 79.08 on SE and ME sub-tasks, respectively. On the other hand, we observe significant decreases on the out-of-domain test set, where the performance drops roughly 16 and 15 points for SE and ME, respectively. These results are reasonable, since domain adaptation is a challenging problem across many NLP tasks. Recent research (Fried et al., 2019) on constituency parsing (a well-studied NLP task) shows that even the latest state-of-the-art parser suffers from large performance drops caused by domain shifting. That parser achieves an F1 score of almost 92% on the Penn Chinese Treebank (Xue et al., 2005), yet its accuracy drops nearly 16 points in the conversational domain. This domain adaptation problem can be much severer for other NLP tasks, such as reading comprehension (Jia and Liang, 2017). Therefore, how to deal with domain adaptation for CASA will be one of important research topics in future.

As an interesting observation, the performance decrease from *ME* to ME_{cross} is not dramatically enlarged on the out-of-domain testset comparing with the in-domain testset. This indicates the robustness of our models and the consistency of our annotations. Intu-

3 `https://huggingface.co/`

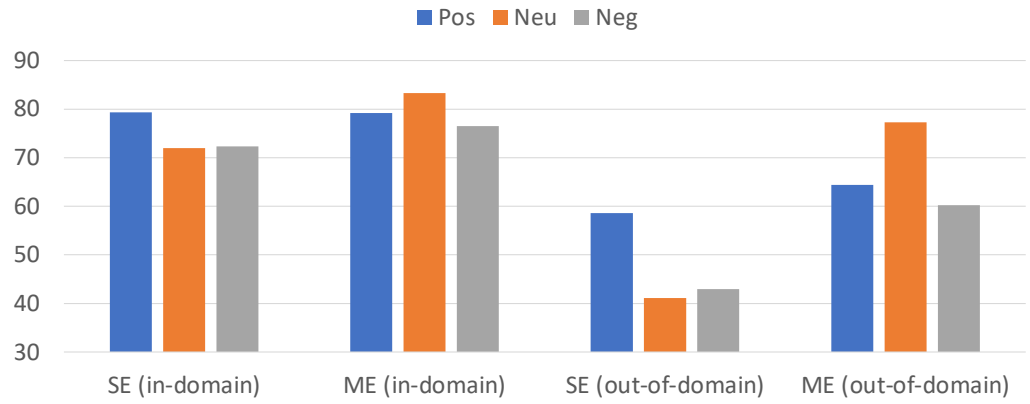

Figure 2: The sentiment and mention extraction performances of *BERT* regarding different polarity values.

Training data	DuConv (in-domain)			NewsDialogue (out-of-domain)		
	ME	ME_{same}	ME_{cross}	ME	ME_{same}	ME_{cross}
All data	79.08	77.59	80.95	63.97	73.95	58.65
Same-turn only	70.02	98.17	34.85	42.97	73.63	26.68
Cross-turn only	68.91	44.85	98.96	45.30	25.12	56.01

Table 6: Mention extraction results (using our *BERT* model) with only same-turn or cross-turn training data, where "ME_{same}" and "ME_{cross}" denote the accuracies for the instances at the same turn and at a different turn, respectively.

itively, cross-turn mention extraction faces additional challenges of long-distance reasoning, including coreference resolution and zero pronoun resolution. Note that the state-of-the-art models for coreference resolution and zero pronoun resolution (Joshi et al., 2019; Yin et al., 2017) only report benchmark performances of 65.9 and 22.7 in Chinese. This indicates that our BERT-based model finetuned with annotated cross-turn data may already have the ability for cross-sentence reasoning to some extent. Likewise, we will focus on how to enhance our model by modeling cross-sentence reasoning.

5.3 Analysis on Different Sentiment Polarities

Figure 2 visualizes the sentiment and mention extraction performances of *BERT* for different sentiment polarities on both the in-domain and out-of-domain test sets. In general, the performances are consistent across different polarities on the in-domain test set. This indicates that the unbalanced label distribution of our CASA dataset has moderate influence on the model performance. However, the variation on the out-of-domain test set is significant. Again, the reason can be that the model underfits the out-of-domain test cases due to the insufficient amount of training data. As the solution, we plan to increase the in-domain training set in future work.

(a)	没(no) 办法(solution) , 大家(people) 都(all) 说(say) 是(is) [**a** 扶不起(perennial) 的 阿斗(losers) **-1**] 。
(b)	在(in) 大连(Dalian) 那个(that) 城市(city) [**a** [**A** 景色(scene) 好(nice) **+1**] **+1**] , [**a** 四季(four seasons) 分明(distinct) **+1**] 。
(c)	朱亚文(Yawen Zhu) [**a** [**A** 太帅(so hansome) **+1**] **+1**] 了 。 [**a** 行走(walking) 的 荷尔蒙(charming) **+1**] ！
(d)	这个(this) 比分(score) 确实(indeed) 让(make) 人(people) [**a** 有点(a little bit) [**A** 气(angry) **-1**] **-1**] 。
(e)	我(I) 认为(think) 他们(they) 比较(relatively) 传统(traditional) , [**a** [**A** 不太(not) 愿意(willing) **-1**] 接受(accept) 新(new) 事物(things) **-1**] 。
(f)	你(you) 是不是(whether or not) [**A** 喜欢(like) **-1**] 马思纯(Sichun Ma) 。
(g)	只要(as long as) 态度(attitude) 认真(serious) , 训练(training) 到位(sufficient) , 国足(Chinese soccer) 成绩(results) [**A** 会(would) 变(become) 好(better) **-1**] 的 。

Table 7: Examples of typical errors in sentiment extraction, where predicted sentiment expressions are surrounded by red brackets with capital letters.

5.4 Same-/Cross-Turn Reasoning for Mention Extraction

Table 6 compares the accuracies of the BERT model on mention extraction when only same-turn or cross-turn training instances are available. When trained only with same-turn data[4], the model reports poor results for cross-turn instances. Similarly, it gets very low accuracies on same-turn instances, when trained only on the cross-turn data. Combining both observations, we can conclude that the ability of cross-turn reasoning cannot be indirectly learned from same-turn instances, and vice versa. The underlying reason is that our model learns an incorrect and biased distribution for making predictions when training on just one type of data.

On the out-of-domain test set, the model trained with all data reports the best performances on same-turn and cross-turn situations, outperforming the models only trained with same-turn or cross-turn instances. It is because the model underfits the out-of-domain data, thus adding extra training instances can generally help.

5.5 Error Analysis

Sentiment extraction Table 7 provides some examples generated by our BERT model from NewsDialogue with typical errors in sentiment extraction. As shown by the three examples from the first group, one major type of errors occurs when our baseline fails to recognize a sentiment expression that is an idiom (e.g. "四季(four seasons) 分明(distinct)") or a new catchword (e.g. "行走(walking) 的 荷尔蒙(charming)"). This is likely because of data sparsity, and they could be alleviated by an external sentiment dictionary with idioms and new catchwords.

The second group demonstrates another type of errors, where the baseline system makes errors for predicting the exact span boundaries. For both cases, the annotators follow the

4 This setting follows the existing work on standard aspect sentiment analysis.

A:	我(I) 平时(usually) 很(very) 喜欢(like) [**a** 研究(study) [**A** 电脑(computer) 手机(mobile) 性能(performance)]] 。
A: B:	[**a** 林宥嘉(Yoga Lin)] 最近(recent) 的 少女("Otomen") 很火(very popular) 啊 。 我(I) 觉得(think) [**A** 他(he) 出(make) 新歌(new songs)] 算是(is) 比较慢(relatively slow) 的 , 产量(productivity) 不高(not high) 。
A: B: ... A:	[**A** 杜兰特(Durante)] 是(is) FMVP 这个(this) 只(only) 看(consider) 得分(score) ! 不看(not) 无不(whether not) 无私(selfless) ! 就是(certainly) , [**a** 库里(Curry)] 确实(indeed) 是(is) 很(very) 伟大(great) , 同时(and) 也(also) 不(not) 忘(forget) 队友(teammates) 。 现在(now) 勇士(Warriors) 既有(has) [**A** 杜兰特(Durante)] 又有(and) 库里(Curry) , 我(I) 还是(still) 喜欢(like) [**a** 库里(Curry)] 。

Table 8: Examples of typical errors for mention extraction. Target sentiment expression is marked with underline, and gold/predicted mentions are surrounded by black/red brackets with lowercased/capital letters.

annotation guideline (Section 3.2) to choose spans with sufficient details, while the baseline system misses some crucial details. Specifically, "有点(a little)" that indicates the extent of "气(angry)" in the first case and "接受(accept) 新(new) 事物(things)" that is the object of "不太(not) 愿意(willing)" are dropped.

The last type of errors, as exhibited in the third group, occurs when an emotional phrase is incorrectly considered as a sentiment expression. This usually happens in an interrogative sentence (e.g. example (f)), where the speaker asks for opinions from others, rather than expressing his/her own ones. Another major type of situations (e.g. example (g)) is in a conditional sentence, where the speaker simply makes a judgement or prediction.

Mention extraction Table 8 contains some frequent errors of our *BERT* model for mention extraction on NewsDialogue. Generally, this model tends to make two types of errors in mention detection. As shown in the first example of Table 8, one type of errors occur when the model inaccurately predicts the boundaries of a mention. As shown in the second and third examples of Table 8, the other type of errors happen when our model predicts a different mention from the the reference. Taking a closer look at the results, we find that our model tends to choose a nearby mention whenever not being confident with other choices. For the second example, the baseline picks "他(he) 出(make) 新歌(new songs)" that is much closer than the correct answer "林宥嘉(Yoga Lin)". Besides, it sometimes tends to pick a popular phrase no matter what the surrounding context is. As seen in the third example, the system picks "杜兰特(Durante)" instead of "库里(Curry)" for both sentiment expressions, even though "库里(Curry)" is much closer to the sentiment expression. As neither of the mentions appears in the training set, the reason for choosing "杜兰特(Durante)" may be that "杜兰特(Durante)" appears more often in the data for training the Chinese BERT.

We conduct human evaluation on 100 dialogues from the out-of-domain test set, counting the mention extraction results as the following situations: (1) correct prediction, (2)

incorrect prediction with erroneous boundaries, (3) incorrect prediction that extracts a coreference pronoun instead, and (4) incorrect prediction that extracts a different mention. Out of the 239 mention extraction cases in the 100 dialogues, the above 4 categories take 71.1%, 2.5%, 7.5% and 18.8%, respectively. This indicates that "incorrectly predicting a different mention" is the current bottleneck for this task, and future work can introduce external knowledge to further distinguish different mentions.

6. The Usefulness of CASA on Knowledge-Driven Dialogue Response Generation

We further evaluate the usefulness of conversational aspect sentiment analysis on knowledge-driven dialogue response generation. In particular, we expect that CASA can help select the relevant parts from massive input knowledge, so that both memory efficiency and performance can be improved.

6.1 Data

We choose NewsDialogue (Wang et al., 2021), which focuses on knowledge-driven dialogue response generation in Chinese. In NewsDialogue, each conversation is created by asking two annotators to discuss a provided news document. The news documents cover a wide range of topics, such as a new technology, a recent basketball game or a movie. The dataset consists 20K conversations, each having around 20 dialogue turns and 416 Chinese characters. This is a good testbed for verifying the effectiveness of CASA, because annotators are asked to conduct conversations *naturally* and *freely* without any restrictions (e.g., fixing the order of discussed topics as in Wu et al. 2019). We separate 272 and 272 dialogue sessions to make the development and test sets, taking the remaining data as the training set.

6.2 Baseline and Model

Similar with Dinan et al. (2019), we adopt a Transformer (Vaswani et al., 2017) model to crate the baselines and our model. One baseline model (i.e., *Transformer* in Table 9) takes only a dialogue history. The other baseline system (i.e., *Transformer w/ Full Doc* in Table 9 with its architecture shown in Figure 3(a)) uses the concatenation of a dialogue history and the corresponding news document as the input before generating a sentence as the response.

Figure 3(b) visualizes our model (i.e., *Transformer* w/ CASA in Figure 3), which uses the same neural architecture except that it only takes some parts of the news document selected according to CASA. In particular, we use our BERT-based CASA model to analyze each dialogue turn, then we choose the entities with positive sentiment in that turn as key phrases, selecting their surrounding contexts from the news document as the knowledge for that turn. In this way, it can be easier for our model to focus on the relevant content in the provided news document than the baselines. The reason for only considering the entities with positive sentiment is that people tend to keep discussing the topics they enjoy talking.

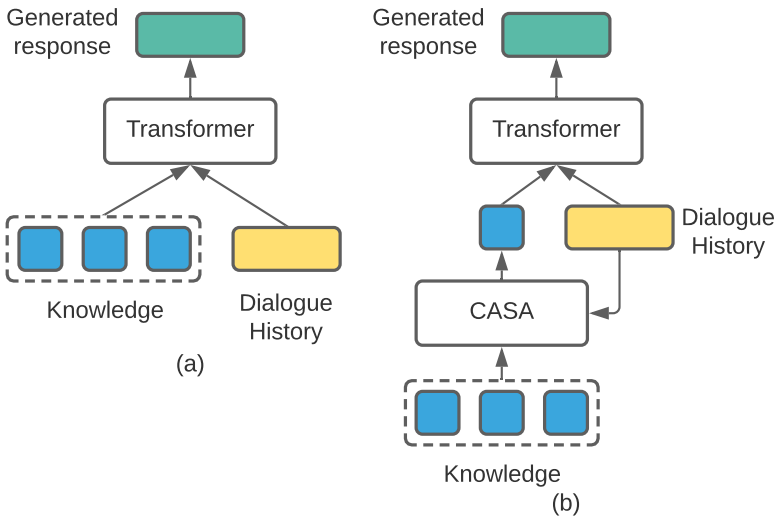

Figure 3: (a) The Transformer baseline and (b) our model using CASA for dialogue response generation.

Model	BLEU-1/2	Distinct-1/2	Avg. KN Len.
Transformer	29.9/17.3	3.6/18.6	0.0
w/ Full Doc	**30.1**/17.3	3.1/16.4	765.5
w/ CASA	29.7/**17.4**	**3.7**/**19.2**	29.1

Table 9: Results on knowledge-driven dialogue response generation, where *Avg. KN Len.* represents the average number of characters in selected knowledge.

6.3 Settings

For both baselines and our model, the encoder and decoder take 4 multi-head self-attention layers, each layer taking 512 hidden units and 8 heads. Adam (Kingma and Ba, 2014) with learning rate 10^{-5} is adopted to train all systems for 50 epochs. The batch size is set to 16 for all systems as well.

6.4 Results

Table 9 compares our model with the baselines under BLEU-1/2 (Papineni et al., 2002) and Distinct-1/2 (Li et al., 2016) scores. We also list the average length of utilized knowledge as demonstrated in column "*Avg. KN Len.*" Using full news documents gives slight increase in BLEU scores, but the diversity of outputs decreases as indicated by the Distinct scores. This is because using full document can introduce noise, as most content within it is not helpful (relevant) to generate the annotated response. Conversely, taking only the selected segments according to CASA improves the diversity regarding Distinct score and shows a comparable BLEU score. More importantly, only 29 Chinese characters on average are selected by CASA, while the full-document baseline uses 765 characters. This shows that CASA can save 96% memory usage for representing relevant knowledge.

7. Conclusion

In this paper, we have proposed the task of *conversational aspect sentiment analysis (CASA)*, which can provide useful symbolic features for general-purpose dialogue understanding. We also annotated a new dataset, as existing datasets on standard aspect sentiment analysis only have a limited number of instances and only focus on a few domains. We also have provided strong baselines using BERT. Our experimental results and deep analysis indicate several potential directions for further improving the performance of CASA. We also demonstrate that CASA can help select knowledge for knowledge-driven dialogue response generation, which improves both memory efficiency and performance.

Future work includes studying how to increase model robustness when applying CASA on different domains. In particular, our plan is to explore the latest achievements of domain-general text understanding (e.g. using public KG and Ontology), leveraging them as rich features to alleviate the common errors we have found in this work.

Acknowledgments

We thank the anonymous reviewers and editors for their insightful comments. Jinsong Su was supported by National Natural Science Foundation of China (No. 62036004), Natural Science Foundation of Fujian Province of China (No.2020J06001), and Youth Innovation Fund of Xiamen (Grant No.3502Z20206059).

References

Asghar, N., Poupart, P., Hoey, J., Jiang, X., and Mou, L. (2018). Affective neural response generation. In *European Conference on Information Retrieval*, pages 154–166. Springer.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Chen, P., Sun, Z., Bing, L., and Yang, W. (2017). Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461.

Colombo, P., Witon, W., Modi, A., Kennedy, J., and Kapadia, M. (2019). Affect-driven dialog generation. 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 3734–3743.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., and Hu, G. (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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2019). Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., and Xu, K. (2014). Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 49–54.

Fried, D., Kitaev, N., and Klein, D. (2019). Cross-domain generalization of neural constituency parsers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 323–330.

Jia, R. and Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Jiang, Q., Chen, L., Xu, R., Ao, X., and Yang, M. (2019). A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6281–6286.

Jiang, S. and de Rijke, M. (2018). Why are sequence-to-sequence models so dull? *EMNLP 2018*, page 81.

Joshi, M., Levy, O., Zettlemoyer, L., and Weld, D. S. (2019). Bert for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5807–5812.

Kessler, J. S. and Nicolov, N. (2009). Targeting sentiment expressions through supervised ranking of linguistic configurations. In *Third International AAAI Conference on Weblogs and Social Media*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kitaev, N., Cao, S., and Klein, D. (2019). Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505.

Kumar, G., Joshi, R., Singh, J., and Yenigalla, P. (2019). Amused: A multi-stream vector representation method for use in natural dialogue. *arXiv preprint arXiv:1912.10160*.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Li, X., Bing, L., Li, P., and Lam, W. (2019). A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6714–6721.

Lian, R., Xie, M., Wang, F., Peng, J., and Wu, H. (2019). Learning to select knowledge for response generation in dialog systems. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5081–5087. AAAI Press.

Liao, X., Xu, H., Sun, L., and Yao, T. (2013). Construction and analysis of the third chinese opinion analysis evaluation (coae2011) corpus. *J. Chin. Inf. Process.*, 1(27):56–63.

Luo, H., Li, T., Liu, B., and Zhang, J. (2019). Doer: Dual cross-shared rnn for aspect term-polarity co-extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 591–601.

Mei, H., Bansal, M., and Walter, M. R. (2017). Coherent dialogue with attention-based language models. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Nan, G., Luo, G., Leng, S., Xiao, Y., and Lu, W. (2021). Speaker-oriented latent structures for dialogue-based relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Nguyen, T. H. and Shirai, K. (2015). Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2514.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *10th International Workshop on Semantic Evaluation (SemEval 2016)*.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Ruder, S., Ghaffari, P., and Breslin, J. G. (2016). A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 999–1005.

Saeidi, M., Bouchard, G., Liakata, M., and Riedel, S. (2016). Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2017). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Shao, Y., Gouws, S., Britz, D., Goldie, A., Strope, B., and Kurzweil, R. (2017). Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205.

Sun, C., Huang, L., and Qiu, X. (2019). Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. 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 380–385.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vinyals, O. and Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Wang, W., Pan, S. J., Dahlmeier, D., and Xiao, X. (2017). Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Wang, X., Li, C., Zhao, J., and Yu, D. (2021). Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14006–14014.

Wang, Y., Huang, M., Zhu, X., and Zhao, L. (2016). Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615.

Wu, W., Guo, Z., Zhou, X., Wu, H., Zhang, X., Lian, R., and Wang, H. (2019). Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804.

Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., and Ma, W.-Y. (2017). Topic aware neural response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Xing, C., Wu, Y., Wu, W., Huang, Y., and Zhou, M. (2018). Hierarchical recurrent attention network for response generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Xu, K., Wu, H., Song, L., Zhang, H., Song, L., and Yu, D. (2021). Conversational semantic role labeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xue, N., Xia, F., Chiou, F.-D., and Palmer, M. (2005). The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207–238.

Yang, H., Zeng, B., Yang, J., Song, Y., and Xu, R. (2019). A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction. *arXiv preprint arXiv:1912.07976*.

Yin, Q., Zhang, Y., Zhang, W., and Liu, T. (2017). Chinese zero pronoun resolution with deep memory network. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1309–1318.

Young, T., Cambria, E., Chaturvedi, I., Zhou, H., Biswas, S., and Huang, M. (2018). Augmenting end-to-end dialogue systems with commonsense knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yu, D., Sun, K., Cardie, C., and Yu, D. (2020). Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Zhao, T., Zhao, R., and Eskenazi, M. (2017). Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.

Zhao, Y., Qin, B., and Liu, T. (2014). Creating a fine-grained corpus for chinese sentiment analysis. *IEEE Intelligent Systems*, 30(1):36–43.

Zhong, P., Wang, D., and Miao, C. (2019). An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7492–7500.