Evaluating Span Extraction in Generative Paradigm: A Reflection on Aspect-Based Sentiment Analysis

Anonymous ACL submission

Abstract

In the era of rapid evolution of generative language models within the realm of natural language processing, there is an imperative call to revisit and reformulate evaluation methodologies, especially in the domain of aspect-based sentiment analysis (ABSA). This paper addresses the emerging challenges introduced by the generative paradigm, which has moderately blurred traditional boundaries between understanding and generation tasks. Building upon prevailing practices in the field, we analyze the advantages and shortcomings associated with the prevalent ABSA evaluation paradigms. Through an in-depth examination, supplemented by illustrative examples, we highlight the intricacies involved in aligning generative outputs with other evaluative metrics, specifically those derived from other tasks, including question answering. While we steer clear of advocating for a singular and definitive metric, our contribution lies in paving the path for a comprehensive guideline tailored for ABSA evaluations in this generative paradigm. In this position paper, we argue that exact match is insufficient as the sole evaluation metric for generative ABSA and that partial match metrics should be used as supplementary measures. We offer guidelines for practitioners navigating ABSA evaluation in the generative era, and further connect this evaluation challenge to annotation quality issues—specifically, the inherent variability and potential errors in span-level annotations.

1 Introduction

Extracting information from user-generated reviews is pivotal and essential in real-world applications. In response to this demand, aspect-based sentiment analysis (ABSA), a task of extracting various aspects of sentiment information from user reviews, has emerged and evolved with various studies (Liu, 2012; Chebolu et al., 2023). The contemporary iterations of ABSA typically refer to the aspect sentiment quad prediction (ASQP) framework (Zhang et al., 2021), encompassing four elements: aspect term, aspect category, opinion term, and sentiment polarity[1]. Fig. 1 shows the two examples that extract one quadruple for each input sentence. In example (A), the aspect and opinion terms, i.e., "*staff*" and "*horrible*", are span factors directly extracted from source sentence(s), whereas aspect category and sentiment polarity, i.e., "Service general" and "Negative", reflect their corresponding categorical attributes. Fundamentally, ABSA is rooted in comprehension; to tackle the problem accurately, a model must grasp both context and underlying intent. However, in the era of generative language models (GLM), researchers encounter a novel realm of challenges in both extracting and classifying these four elements. This

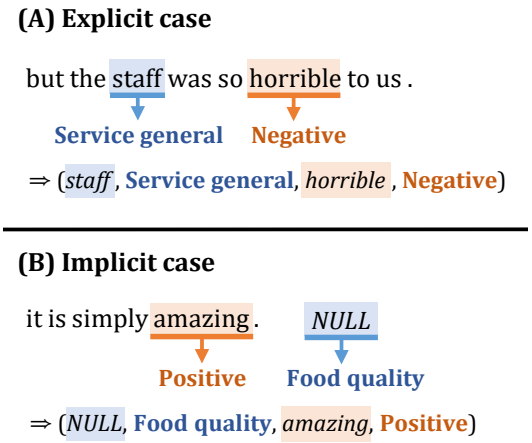

Figure 1: Aspect sentiment quad prediction (ASQP) examples from ACOS (Cai et al., 2021) rest16 dataset. Each quadruple is extracted from a given sentence in the order of (aspect term a, aspect category c, opinion term o, sentiment polarity s). Example (A) is an explicit case where the mentions of aspect and opinion terms are described in the given sentence, while (B) is an implicit one where the aspect term is not found in the sentence.

[1] In this paper, our discussions operate under the default ASQP configuration that predicts all four elements.

challenge has been amplified by the recent deployment of large language models (LLMs) such as ChatGPT and LLaMA for ABSA tasks (Zhang et al., 2024; Bai et al., 2024), where generated outputs frequently contain semantically valid but surface-different spans compared to ground truth annotations.

Adding to the complexity, recent ABSA methodologies, such as ACOS dataset (Cai et al., 2021), have incorporated a "*NULL*" annotation to address instances where span tagging is not applicable, echoing the challenges faced with unanswerable scenarios in question-answering (QA) tasks (Rajpurkar et al., 2018). As shown in example (B) of Fig. 1, the aspect term is tagged as "*NULL*" because the term is not represented as an entity explicitly, unlike the opinion term "*amazing*". Moreover, a broader spectrum of domains is covered with the latest expansions in the MEMD-ABSA dataset (Cai et al., 2023), introducing new domains of books, clothing, and hotel, along with the explored areas of laptop and restaurant. This inclusivity presents a diverse range of tendencies across domains, and the potential aspect and opinion candidates similarly exhibit extensive variability.

Despite the ongoing research in models and datasets to perform ABSA tasks, there has been less discussion on the evaluation scheme of their outcomes. In fact, evaluating the output produced by the ABSA model(s), particularly implemented using the extract-classify-ACOS methodology (Cai et al., 2021), poses intricate hurdles, primarily because it encompasses two concurrent abilities of extraction and classification, each demanding simultaneous attention. As shown in Fig. 1, the ABSA task involves extraction and classification processes, leading to the formation of pairs: (aspect term, aspect category) and (opinion term, sentiment polarity), where the former requires extraction and the latter necessitates categorizing ability. On the surface, this dichotomy, pairing the aspect term with the aspect category and the opinion term with the sentiment polarity, seems logically structured. However, this division doesn't necessarily indicate that the decision-making process for one attribute is wholly independent of the other since the spans of aspects and opinions terms influence the decision process of each other and sometimes they can intersect.

Assessing individual instances in ABSA is further complicated by the predictions of multiple quadruples against several ground truth quadruples. This raises two salient questions: *how should each instance be assessed when there's no perfect alignment of sequences? How should one measure the similarity score between predictions and ground truths?* Furthermore, the evolution and ubiquitous adoption of generative models (Radford et al., 2018; Brown et al., 2020; Raffel et al., 2020) introduces another layer of complexity. With these models, there's a greater potential for diverse answers. Thus, the pertinent question becomes: *how should such diverse responses be evaluated? Should they be perceived and assessed under the same lens as the traditional extract-and-classify scenarios?*

To resolve the aforementioned complexities, this position paper discusses how diverse predictions and ground truths are evaluated and how we can evaluate the outputs of the GLM era. In Section 2, we simply summarize the notation and the distinct subtasks inherent to the ABSA task. Note that we only deal with the essential components to discuss the evaluation-related topics in the context of this position paper. In Section 3, we survey how ABSA has been conducted prior to the generative era and explore the ongoing trends, particularly emphasizing inferential attempts facilitated by pretrained and generative language models. Followingly, in Section 4, we take a look at conventional ABSA evaluation schemes, including exact match and F1 metrics, while also shedding light on prevalent similarity measures used to evaluate term and expression correspondence. Lastly, in Section 5, we compare evaluation schemes from various perspectives along with diverse examples and offer a suggestion on the future direction of ABSA evaluation. The primary goal of our manuscript is to spotlight challenges with the adoption of conventional ABSA evaluation in the emergent paradigm of generative models and to further establish a guidelight for assessing the model output with multiple span extraction problems.

2 Background

2.1 Four elements of ABSA

Sentiment analysis is the study that analyzes individuals' sentiments and perceptions towards entities and their attributes (Liu, 2012). In the ABSA task, the opinion term (o) and sentiment polarity (s) belong to the prior sentiment-related elements, while the aspect term (a) and aspect category (c) go to the entity-related elements. The definition of

these four elements can be described as follows:

- **Aspect term** (a) refers to a target entity or object about which an opinion is expressed, usually represented by a word or phrase, but it can occasionally be implicit, as shown in example (B) of Fig. 1.
- **Opinion term** (o) conveys a sentiment or viewpoint about the aforementioned aspect. This is usually described in words or phrases, but it can be implicit.
- **Aspect category** (c) classifies the aspect into a specific class selected from a predefined category set.
- **Sentiment polarity** (s) is the sentiment the author has for the aspect, which is usually revealed by the opinion term o and divided into three classes: *positive, neutral*, and *negative*.

2.2 Subtasks of ABSA

Task	Output
Aspect term extraction (ATE)	a
Aspect-opinion pair extraction (AOPE)	a, o
Aspect-sentiment pair extraction (ASPE)	a, s
Aspect-sentiment triplet extraction (ASTE)	a, o, s
Aspect-category-sentiment detection (ACSD)	a, c, s
Aspect-sentiment quad prediction (ASQP)	a, c, o, s

Table 1: Target output elements for ABSA tasks. The names of the tasks follow Zhang et al. (2022).

As shown in Table 1, several inference types can be identified when discussing ABSA, with the more prominent ones being: aspect term extraction (ATE) (Hu and Liu, 2004), aspect-opinion pair extraction (AOPE) (Fan et al., 2019), aspect-sentiment pair extraction (ASPE), aspect-sentiment triplet extraction (ASTE) (Peng et al., 2020), aspect-category-sentiment detection (ASCD), and aspect-sentiment quad prediction (ASQP). As a representative dataset, SemEval datasets (Pontiki et al., 2014, 2015, 2016) introduces a wealth of resources applicable to ATE, ASPE, and ACSD tasks. In a recent study, the ASQP (Zhang et al., 2021) task, which considers all four elements, has gained attention. However, because all four factors should be extracted from a given sentence(s), there may be implicit cases where aspects or opinion terms do not appear explicitly in the input sentence(s). To encompass such implicit cases, datasets that additionally include "*NULL*" spans, e.g., ACOS (Cai et al., 2021) and MEMD-ABSA (Cai et al., 2023), have been introduced. Note that datasets for the ASQP task can be utilized in all the subtasks mentioned earlier. In the scope of this position paper, we primarily mention the essential subtasks and benchmarks aligned with the ASQP task. Comprehensive surveys encapsulating a broader collection of ABSA models and benchmarks can be found in works by Zhang et al. (2022) and Chebolu et al. (2023).

3 ABSA Models in Paradigm Shift

3.1 Inference De Facto

Traditional inference process in ABSA artfully combines a variety of techniques (Chebolu et al., 2023). It blends named entity recognition (NER)-style span extraction, the linking of aspect and opinion terms, as depicted by the highlighted spans in Fig. 1, and the categorical decision-making process on both category and sentiment, as illustrated by the arrows in Fig. 1. Also, this process can either incorporate binary/ternary (sentiment) or multiple (category) labels in nature. Prior to the emergence of large-scale models, the tasks of extraction and prediction have been processed in distinctly separate stages. This bifurcation boosted methodologies akin to NER and categorization (Saias, 2015; Phan and Ogunbona, 2020).

The advent and widespread utilization of BERT-style (Devlin et al., 2019) pretrained language models (PLMs) found their synergistic alignment with the aforementioned inferential framework. This harmony is exemplified in the extract-classify method, elucidated in the foundational ACOS paper (Cai et al., 2021). This approach merges span-tagging extraction with subsequent prediction layers —- all integrated within a single model. Across these methodologies, there's a consistent theme: aspects and opinions are systematically extracted from the provided input, while categories and sentiments are predicted based on a predefined set of classes.

3.2 Recent Attempts with GLM

Employing generative models for ABSA inference has garnered significant attention, primarily driven by the advent of landmark GLMs like GPT-3 (Brown et al., 2020) and sequence-to-sequence (Seq2seq) styled language generation techniques such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). Among them, T5 has gained

paramount prominence. To bolster the accuracy of these generative models and curb tendencies for hallucination, researchers have begun experimenting with constrained decoding (De Cao et al., 2020) during the generation phase. This technique aids the model in selecting tokens directly from the input sentence or the category and sentiment sets. With the constrained decoding, T5-centric methodologies (Zhang et al., 2021; Mao et al., 2022; Bao et al., 2022; Hu et al., 2022; Gou et al., 2023) reign supreme in this domain. This typically involves feeding the model with instance-quadruple pairs for training and testing. The ultimate objective is to fine-tune T5 into a model adept at decoding the appropriate set of quadruples derived from a single instance.

The capabilities of GLMs have been advancing at a remarkable pace, and their applicability has broadened remarkably. Initially, GLMs found their primary niche in inherently *generative* tasks, such as machine translation, story generation, or dialogue management. However, the unveiling of language modeling techniques that enable both comprehensive understanding and generative faculties, i.e., InstructGPT (Ouyang et al., 2022), and its latecomers, brought about a novel approach such as in-context learning. They were also employed to proceed with categorical data; thus, metrics like accuracy and F1 score became instrumental in gauging their efficacy.

It is important to note that metrics for evaluating existing models have been performed separately for classification, e.g., precision, recall, and F1 scores, and generation, e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019). However, with the advent of GLMs, decoder-based models are actively used for classification tasks (Min et al., 2022a,b; Yoo et al., 2022), in addition to generation tasks where decoder-only or encoder-decoder models have traditionally been used. In ABSA, the transition of the backbone model from BERT to T5 also echoes this trend.

3.3 ABSA with Large Language Models

The landscape of ABSA has been further transformed by LLMs. Zhang et al. (2024) conducted a comprehensive evaluation across 13 sentiment analysis tasks and 26 datasets, finding that while LLMs lag behind fine-tuned small models on structured ABSA tasks, they significantly outperform in few-shot settings. Instruction-tuned approaches such as InstructABSA (Scaria et al., 2024) have demonstrated competitive performance using a fraction of training data, while chain-of-thought reasoning strategies have been applied to improve ASQP quality (Kim et al., 2024). Critically for our discussion, these LLM-based approaches expose a fundamental limitation of exact match evaluation: LLMs frequently generate semantically correct predictions that differ in surface form from ground truth annotations.

4 Evaluation Schemes

4.1 ABSA Evaluation Schemes

As with the development of ABSA inference approaches, the evaluation landscape of ABSA has also been rich with distinct metrics tailored to the specific components of the analysis. For the categorical attributes, namely category and sentiment, widely accepted metrics such as accuracy and F1 score are employed. This decision is reasonable, given that the pool of candidates for these attributes is finite. However, it's worth noting that there exists an inherent disparity in the number of candidates across these attributes; while categories generally offer a broader selection compared to the sentiments, which are typically capped at three, albeit with the added layer of subjectivity.

When it comes to evaluating aspects and opinions, the exact match metric is the go-to (Cai et al., 2021; Gou et al., 2023). However, this metric is notably harsh for span outputs, primarily owing to the decision schemes associated with aspects and opinions, especially from the viewpoint of manual data construction. These schemes are often contingent on the domain of the corpus and the type of sentences. Complicating matters further is the intrinsic challenge of distinguishing between aspects and opinions, and it is difficult to establish rules of thumb for consistently dissecting them in certain circumstances. For instance, if we have three sentences:

(1a) "The *dinner* was so *expensive*."

(1b) "The *price* was *high* in the dinner time."

(1c) "The *dinner price* was quite *high*."

all the categories indicate the "Price" without doubt, but deciding the aspect and opinion for (1c) would be somewhat obscure given (1a) and (1b), which may invoke inconsistency in the annotation. Also, "*expensive*" itself is the term that implies the infor-

mation on pricing, which adds a challenge. Example (1b) also illustrates the challenge of *discontinuous* aspect spans: the full intended aspect may be " *price in the dinner time* ," with " *price* " as a central component and " *in the dinner time* " as an auxiliary modifier (Cai et al., 2023). Whether to annotate the full discontinuous span or only the head noun is an annotation guideline decision that directly affects evaluation outcomes, as exact match would penalize any mismatch. This is yet another case where annotation inconsistency can be conflated with model prediction error. If these kinds of discrepancies and challenges occur simultaneously in training and test sets, it would result in an inadvertent disadvantage of potentially appropriate predictions if the exact match is the only metric utilized.

For those seeking a more lenient evaluation metric for aspects and opinions, partial match (Ku et al., 2008; Li et al., 2022) schemes such as word-level F1 score can be an attractive alternative. Specifying the score into word or token level can provide a more delicate assessment than the exact match, which demands perfection. It allocates scores even when there's a slight variance in expression, accommodating those instances that the exact match metric would deem incorrect. On the other hand, given that the correct answer can be extracted in diverse forms, e.g., whether the GTs contain adverbs or not, we can refer to the dataset construction and evaluation methodology of machine reading comprehension (MRC) tasks. In the case of SQUAD 2.0 (Rajpurkar et al., 2018) and KLUE-MRC (Park et al., 2021), the MRC task is performed to find the domain of a question in a given context, and several candidates are prepared according to the inclusion of articles such as 'the', and the model prediction is considered successful if the output generated by the model exists in the candidates. In ABSA task, it can be another alternative to consider the additional GT candidates for aspect and opinion terms.

One crucial part of ABSA inference is that it capacitates multiple quadruple predictions and multiple ground truth answers (GTs) for a single instance. Presently, precision and recall are mainly investigated in the evaluation; that is, the evaluation process incorporates counting *effective* predictions and all GTs that are correctly predicted. However, the evaluation doesn't necessarily factor in scenarios where the predictions do not necessarily exact match with GT but are informative enough to be assessed as an effective prediction. In such circumstances, both the exact match and partial metrics can impose a rather rigorous evaluation criterion.

4.2 Studies on NLG Evaluation

With the ascendancy of generative models as the standard for various downstream tasks, the shift towards incorporating natural language generation (NLG) evaluation schemes in lieu of traditional methodologies is both palpable and pragmatic. A substantial segment gravitates towards similarity metrics (Sai et al., 2022). However, while there exists a plethora of sentence-level metrics, such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004), these don't seamlessly align with the task at hand, since neither the predictions nor the GTs are not necessarily sentence-level expressions. Particularly at the level of entity or phrase, the word-level F1 score stands out in the representative area of question answering (Rajpurkar et al., 2016), furnishing both precision and recall metrics for generated outputs.

Another avenue worth exploring is PLM-based approaches, such as BERTScore (Zhang et al., 2019). However, a caveat accompanies this strategy: it's profoundly sensitive to domain-specific influences, which might skew evaluations based on the domain's inherent characteristics, that may have discrepancy with the property of the pretraing corpus. The sensitivity of PLM-based metrics to domain characteristics warrants careful consideration; for instance, the pre-training corpus composition may favor certain domains over others, potentially skewing similarity scores in domain-dependent ways.

Concerning the above deliberations, the consensus tilts towards adopting a variety phrase-level semantic similarity metrics that can be applied to aspect and opinion. Here, the partial match schemes discussed above can be again considered as a frontrunner. Its strength lies in calculating precision and recall based on the overall similarity of expression, which can compensate for the harshness of the exact match, and sometimes cover the cases when there is no overlap but the meaning is shared.

5 Discussion

5.1 Comparison of Evaluation Schemes

Considering all evaluation schemes that apply to current ABSA literature, we can come up with the following main topics of discussion.

5.1.1 Using Exact Match vs. Partial Match

Transitioning into a generative paradigm introduces challenges in ensuring that generated terms align precisely with the intended GT spans. While strategies like constrained decoding can help an accurate exact match, it necessitates a reflection on the choice of evaluation metrics: should one adhere to the strict exact match criterion or explore more lenient alternatives like word-level F1 score or other similarity measures? For instance, in (1c), given that the answer is "*dinner price*", though exact match maximizes the utility of precise prediction of span extraction, devaluating predictions such as "*dinner*" or "*price*" without consideration on the approximation of semantics might be harsh concerning the potential utility of the model.

Additionally, a subtle difference exists between various partial match metrics, which have pros and cons depending on domain, sentence/answer type, and dissecting schemes, e.g., whether to adopt whitespace, morphological decomposition, or other tokenization methodologies. Also, it is important whether to take into account the order of words in evaluation or not, which may lead to the superiority of metrics such as longest common substring (LCS) (Li et al., 2022).

5.1.2 Total vs. Element-Wise Evaluation

When assessing quadruple predictions with GTs, the word-level F1 score doesn't account for the variances across individual elements. Given the distinct objectives and evaluative natures of different attributes, the element-wise evaluation might offer a more comprehensive assessment of model performance than the total one. This approach ensures that each attribute's unique characteristics and challenges are taken into consideration during the evaluation. For instance, in the same generative paradigm, evaluation on category and sentiment should better be an exact match considering that the predefined set of candidates is provided, while a partial answer should be tolerated for aspect and opinion terms given that the set of candidate sequences varies from input sentences.

In sum, the quadruple-level assessment can shed light on the *total* performance of the model, given that the score is added only if the whole inference is correct for all elements. However, it is questionable that aspect/category and opinion/sentiment are evaluated with the same criteria just because they are yielded as an output of a generative model.

	Exact match	Partial match
Total	$s \in \{0, 1\}$	$s \in [0, 1]$
Element -wise	$s = \{s_i\}_{i=1}^{N_e}$ $where\ s_i \in \{0, 1\}$	$s = \{s_i\}_{i=1}^{N_e}$ $where\ s_i \in [0, 1]$

Table 2: Score formulation for four evaluation schemes. s is the score, N_e is the number of elements to be considered, where N_e is 4 for ASQP task that $\{1, 2, 3, 4\}$ corresponds with (a, c, o, s).

5.1.3 Case Study

To compare the evaluation metrics with a detailed example, we sample a data instance from ACOS-laptop16 dataset. Assume that we have a generative model, e.g., T5, and the model returns two quadruples for a single input sentence. The input, GT, and generated quadruples can be described as follows, where the extracted quadruples follow the order of (a, c, o, s).

x = "*key* presses are too *stiff* to press ."

g = (*key*, Keyboard usability, *stiff*, Negative)

p_1 = (*key*, Keyboard usability, *too stiff*, Negative)

p_2 = (*key presses*, Keyboard usability, *stiff*, Negative)

x is an input sentence, g is a GT quadruple of x, and p_1 and p_2 are the predictions of a model. Note that p_1 and p_2 have slightly different o and a mentions compared to g, respectively.

Now, let $f_n(g, p_i)$ be a score function of g and p_i that corresponds to four cases of Table 2. The scoring metric is accuracy here. Also, as shown in Table 2, the total scores are scalar-valued, while the element-wise scores can be represented as sequences of length 4 in this case study.

For total and exact match case (function f_1), the score of p_1 and p_2 are both zero, i.e., $f_1(g, p_i) = 0$, because the aspect and opinion terms are not exactly the same with the GT (g). For total and partial match case (function f_2), the score of p_1 and p_2 are both $5/6$, i.e., $f_2(g, p_i) = 0.83$; note that for partial match scores, we count the common (whitespace-split) words between the GT and prediction. While the total-exact match score concentrates on the wrongly predicted span, the total-partial match score shows the overall accuracy of the generated prediction.

On the other hand, the element-wise scores in this ASQP case can be represented as a quadruple following the same order of (a, c, o, s). The element-wise exact match case (function f_3) of

p_1 is $(1., 1., 0., 1.)$ and that of p_2 is $(0., 1., 1., 1.)$, highlighting the element where the model wrongly generated. However, these scores also reveal the harshness as a metric of the exact match. Lastly, the element-wise partial match case (function f_4) of p_1 is $(1., 1., 0.5, 1.)$ and that of p_2 is $(0.5, 1., 1., 1.)$, showing the potential of predictions that would have been underestimated concerning other metrics.

To add complexity to the utility of partial matching metrics, the opinion terms with negating expressions (e.g., *"no", "not", "less"*) can be considered as a challenging real-world example not handled in this case study. Imagine that a generative model (probably without constrained decoding) generates another prediction:

p_3 = (*key*, Keyboard usability, *not stiff*, Neutral)

by inserting "*not*" in the opinion term. In this case, the score of element-wise partial match(f_4) is $(1., 1., 0.5, 0.)$. However, one may find it questionable that the opinion score of p_3, namely $f_4(g, p_3)_3$, is the same as that of p_1, namely $f_4(g, p_1)_3$, which displays the opinion term semantically similar to the GT. In this case, NLG metrics that consider semantic similarity would be an auxiliary measure that can penalize and filter out the predictions that provide contrary meanings and distort the evaluation. Simply in this example where the meaning of the opinion term of p_3 significantly contradicts that of GT, an NLG metric can assign 0, i.e., $f_4(g, p_3) = (1., 1., 0., 0.)$, for the sake of reasonable score assignment.

Recent empirical studies have quantified the magnitude of the evaluation gap illustrated above. Yang et al. (2025) introduced the Zoom In-N-Out framework, which automatically generates alternative valid aspect and opinion terms using LLMs to expand evaluation sets, and demonstrated up to 10 percentage-point improvements in correlation with human judgments (measured by Kendall's τ) on ASQP and ACOS benchmarks. Similarly, Hua et al. (2025) proposed FTS-OBP, combining ROUGE-L-based flexible matching for span components with exact match for categorical components and optimal bipartite pairing for tuple alignment, reporting a mean +0.16 macro-F1 increase over traditional exact match. These findings provide concrete empirical support for the position that exact match alone is insufficient for generative ABSA evaluation.

5.2 Future Suggestion for Generative Paradigm

Along with the four perspectives discussed above, we provide our viewpoints on each topic.

Partial match metrics should be supportively used. The exact match metric, while fitting in pre-generative paradigms where span tagging and categorization were considered subsequent processes, might not be the sole reliable yardstick in a generative setting. Introducing metrics like word-level F1, LCS (Li et al., 2022), or edit distance could serve as indicators of the model's partial success, tempering the strict nature of the exact match. Ideally, a combination of these metrics would provide a more rounded perspective, showcasing both the precision and potential of predictions. We can also apply this idea to the element-wise evaluation that is to be discussed followingly.

Quadruple-level aggregation can be comprehensive, but element-wise evaluation can highlight the characteristics of each element. Recognizing the distinct natures of each attribute is essential in evaluations. Especially the contrasts between categorical and span-based inferences demand acknowledgment. Even if the generative paradigm renders categorical outputs as textual representations rather than logits, their evaluation would be more suited to a categorical framework. For instance, a subtle issue arises when partial match is applied to categorical attributes. Some aspect categories in ABSA follow a two-level hierarchy, e.g., "Food#Quality." If partial match is applied at the token level, a prediction of "Food#Price" would receive a 50% score against the GT "Food#Quality" due to the shared "Food" component, despite representing a fundamentally different category. We argue that categorical attributes should retain exact match evaluation, as partial token overlap in category labels does not carry the same semantic proximity as partial overlap in span expressions. Sentiment polarity, consisting of only three classes, similarly should not benefit from partial match.

Exact quadruple match should accommodate partial match metrics in assessing prediction-GT pairs. Evaluating entire quadruples collectively might obscure the nuances and variations that exist between individual predictions or GTs. Precision and recall, in this context, should be viewed through the lens of the prediction-GT similarity within individual quadruple pairs rather than

searching for the existence of exactly the same output. Specifically, we can assume an adjusted scoring scheme based on a holistic comparison of entire predictions with their corresponding GTs and vice versa. For instance, if three predictions correspond closely to just one out of two GT quadruples, scores pertaining to the three prediction-GT pairs could be consolidated. Simultaneously, there should be a deduction in the aggregate score to account for the overlooked gold quadruple. As an example, we can think of the following concept of a system:

Let $P = \{p_1, p_2, p_3, p_4\}$ be a set of predictions for a single instance whose ground truth is $G = \{g_1, g_2, g_3\}$. In de facto exact match-based evaluation, precision and recall would be calculated concerning if $p_i \in G$ for all i and if $g_j \in P$ for all j. However, assuming that $\{p_1, p_2\}$ were exact match or close guess to g_1 (in terms of a similarity metric), $\{p_3, p_4\}$ to g_2, and none of them were relevant to g_3, as following quadruples, which follow the order of (a, c, o, s):

p_1 = (*dinner*, Price, *so expensive*, Negative)
p_2 = (*dinner price*, Price, *so high*, Negative)
p_3 = (*beverage*, Food quality, *too cold*, Negative)
p_4 = (*beverage*, Drink, *cold*, Positive)

g_1 = (*dinner*, Price, *so expensive*, Negative)
g_2 = (*beverage*, Drink, *too cold*, Negative)
g_3 = (*lamb steak*, Food quality, *awesome*, Positive)

set aside from the harsh exact match metric which may give the precision of 0.25 and the recall of 0.33, we can think of weight averaging all relevant similarity scores regarding $(p_1, g_1), (p_2, g_1)$ and $(p_3, g_2), (p_4, g_2)$ as a correspondence to $\{g_1, g_2\}$ and penalize the whole score for not even getting close to g_3. Note that this concept is just a recommendation; though it does not strictly follow the conventional exact match scheme, it accommodates partial exact match metrics and also allows an element-wise analysis.

5.3 From Evaluation to Annotation

Our discussion on evaluation metrics is intimately connected to annotation quality. Inconsistencies in span boundary decisions—such as whether to include modifiers or articles in aspect/opinion annotations—are not merely stylistic choices but can constitute annotation errors (Klie et al., 2023). For ABSA specifically, the reliance of approximately 78% of research on a small set of SemEval benchmarks (Chebolu et al., 2023) raises concerns about the compounding effect of undetected annotation errors. Systematic annotation error detection study has been rarely applied to standard ABSA benchmarks using modern AED methods, and we advocate for such an investigation as a necessary complement to the evaluation metric improvements discussed in this paper: flexible metrics can mitigate the impact of annotation variability, but identifying and correcting objective errors in ground truth data is equally important for reliable ABSA evaluation.

In this regard, approaches such as the aforementioned Zoom In-N-Out framework (Yang et al., 2025) suggest creating an alternative evaluation set with more flexible answer candidates, thereby extending the original annotations to cover expressions that cannot be directly generated by continuous span extraction. Though ABSA datasets are primarily released with a single ground truth, these attempts pave the way for accommodating imaginable semantic variants and fairly evaluating the unfortunately missed ones. Our proposed guideline for the generative paradigm would provide useful references not only in the evaluation phase, but also in the annotation phase for alternative ground truth, where a single 'gold' answer can serve as a seed for reasonable extensions.

6 Conclusion

In this paper, we have examined the evaluation landscape of aspect-based sentiment analysis in the context of the ongoing shift toward generative models. Our central argument is that exact match, while appropriate for the pre-generative paradigm of span tagging and categorical prediction, is insufficient as the sole evaluation metric for generative ABSA. We advocate for element-wise evaluation that applies exact match to categorical attributes (category and sentiment) and partial match metrics to span-based attributes (aspect and opinion), supplemented by semantic similarity measures where appropriate. We will further connect this evaluation challenge to broader concerns about annotation quality: the inherent variability in span-level annotations means that some "errors" flagged by exact match may reflect legitimate annotation alternatives rather than genuine model failures. We hope this study serves as a catalyst for developing standardized, flexible annotation and evaluation protocols for ABSA in the generative era.

Limitation

Our focus in this discussion has been predominantly on refining the metrics for ACOS predictions, particularly considering the inclusion of the *NULL* entity as proposed in the original paper. However, it's essential to acknowledge the limitations of our approach. In this discourse, we've exclusively explored cases pertaining to explicit aspects and opinions, deliberately sidelining instances with implicit terms. Such cases, rich in their inherent complexities, are earmarked for future exploration and analysis.

Additionally, it's worth noting that our intention here isn't to prescribe a definitive metric. Rather than pinpointing an optimal direction for a specific objective, our endeavor has been to shed light on the advantages and disadvantages of various methodologies. Our aim remains to offer a balanced perspective, equipping practitioners with insights that can guide their evaluative processes. Thus, while we have drawn on empirical findings from recent studies to support our arguments, we have not conducted new experiments that supports the validity of a specific metric. An empirical comparison of exact match, partial match, and semantic similarity metrics across multiple ABSA models constitute complementary evidence for our position.

Acknowledgments

To be added.

References

Yinhao Bai, Zhixin Han, Yuhua Zhao, Hang Gao, Zhuowei Zhang, Xunzhi Wang, and Mengting Hu. 2024. Is compound aspect-based sentiment analysis addressed by LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7836–7861, Miami, Florida, USA. Association for Computational Linguistics.

Xiaoyi Bao, Wang Zhongqing, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4044–4050. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hongjie Cai, Nan Song, Zengzhi Wang, Qiming Xie, Qiankun Zhao, Ke Li, Siwei Wu, Shijie Liu, Jianfei Yu, and Rui Xia. 2023. Memd-absa: A multi-element multi-domain dataset for aspect-based sentiment analysis. *arXiv preprint arXiv:2306.16956*.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2023. Survey of aspect-based sentiment analysis datasets. In *2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

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 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. 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 2509–2518, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth*

ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177.

Yan Cathy Hua, Paul Denny, Jörg Wicker, and Katerina Taškova. 2025. Data-efficient adaptation and a novel evaluation method for aspect-based sentiment analysis. *arXiv preprint arXiv:2511.03034*.

Jieyong Kim, Ryang Heo, Yongsik Seo, SeongKu Kang, Jinyoung Yeo, and Dongha Lee. 2024. Self-consistent reasoning-based aspect-sentiment quad prediction with extract-then-assign strategy. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7295–7303, Bangkok, Thailand. Association for Computational Linguistics.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198.

Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. 2008. Question analysis and answer passage retrieval for opinion question answering systems. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 13, Number 3, September 2008: Special Issue on Selected Papers from ROCLING XIX*, pages 307–326.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022. MultiSpanQA: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1260, Seattle, United States. Association for Computational Linguistics.

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

Bing Liu. 2012. *Sentiment analysis and opinion mining*. Synthesis lectures on human language technologies. 5 vol.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2Path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2215–2225, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, and 12 others. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8600–8607.

Minh Hieu Phan and Philip O. Ogunbona. 2020. Modelling context and syntactical features for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3220, Online. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 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, Austin, Texas. Association for Computational Linguistics.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

José Saias. 2015. Sentiue: Target and aspect based sentiment analysis in SemEval-2015 task 12. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 767–771, Denver, Colorado. Association for Computational Linguistics.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Sawant, Swaroop Mishra, and Chitta Baral. 2024. InstructABSA: Instruction learning for aspect based sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 720–736, Mexico City, Mexico. Association for Computational Linguistics.

Soyoung Yang, Hojun Cho, Jiyoung Lee, Sohee Yoon, Edward Choi, Jaegul Choo, and Won Ik Cho. 2025. Single ground truth is not enough: Adding flexibility to aspect-based sentiment analysis evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12071–12096, Albuquerque, New Mexico. Association for Computational Linguistics.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.