Low-resource Entity Set Expansion: A Comprehensive Study on User-generated Text

Yutong Shao[1*], **Nikita Bhutani**[2†], **Sajjadur Rahman**[2†], **Estevam Hruschka**[2]
[1]University of California, San Diego [2]Megagon Labs
yshao@eng.ucsd.edu, {nikita,sajjadur,estevam}@megagon.ai

Abstract

Entity set expansion (ESE) aims at obtaining a more complete set of entities given a textual corpus and a seed set of entities of a concept. Although it is a critical task in many NLP applications, existing benchmarks are limited to well-formed text (e.g., Wikipedia) and well-defined concepts (e.g., countries and diseases). Furthermore, only a small number of predictions are evaluated compared to the actual size of an entity set. A rigorous assessment of ESE methods warrants more comprehensive benchmarks and evaluation. In this paper, we consider user-generated text to understand the generalizability of ESE methods. We develop new benchmarks and propose more rigorous evaluation metrics for assessing performance of ESE methods. Additionally, we identify phenomena such as non-named entities, multifaceted entities, vague concepts that are more prevalent in user-generated text than well-formed text, and use them to profile ESE methods. We observe that the strong performance of state-of-the-art ESE methods does not generalize well to user-generated text. We conduct comprehensive empirical analysis and draw insights from the findings.

1 Introduction

Entities are integral to applications that require understanding natural language text such as semantic search (Inan et al., 2021; Lashkari et al., 2019), question answering (Chandrasekaran et al., 2020; Cheng and Erk, 2020) and knowledge base construction (Goel et al., 2021; Al-Moslmi et al., 2020). To this end, entity set expansion (ESE) is a crucial task that uses a textual corpus to enhance a set of seed entities (e.g., 'mini bar', 'tv unit') with new entities (e.g., 'coffee', 'clock') that belong to the same semantic concept (e.g., room features).

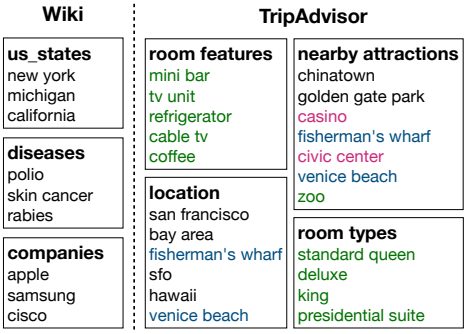


Figure 1: Example concepts and entities from Wiki vs. Tripadvisor. We highlight example multifaceted entities in blue, non-named entities in green and vague entities in magenta.

Since training data in new domains is scarce, many existing ESE methods expand a small seed set by learning to rank new entity candidates with limited supervision. Broadly speaking, there are two types of such low-resource ESE methods: (a) corpus-based methods (Shen et al., 2018; Huang et al., 2020a; Yu et al., 2019a) that bootstrap the seed set using contextual features and patterns, and (b) language model-based methods (Zhang et al., 2020a) that probe a pre-trained language model with prompts to rank the entity candidates.

Despite the recent progress, reported success of ESE methods is largely limited to benchmarks focusing on named entities (e.g., countries, diseases) and well-written text such as Wikipedia. Furthermore, the evaluation is limited to top 10-50 predictions regardless of the actual size of the entity set. As a result, it is unclear whether the reported effectiveness of ESE methods is conditional to datasets, domains, and/or evaluation methods.

In this paper, we conduct a comprehensive study to investigate the generalizability of ESE methods in low-resource settings. Specifically, we focus on

[*]Work done during an internship at Megagon Labs.
[†]Equal author contribution.

domains with user-generated text such as reviews. User-generated text data is abundant and is largely unlabeled. Enabling NLP applications including semantic search and question answering (Li et al., 2019; Bhutani et al., 2020; Dai and Song, 2019) over user-generated text requires entities mined from these largely unlabeled data. Furthermore, user-generated text has distinctive characteristics than well-written text, making it appropriate for this study. Due to lack of benchmarks on user-generated text, we create new benchmarks from three domains – *hotels*, *restaurants* and *jobs*.

We found that these benchmarks exhibit characteristics (illustrated in Figure 1) distinct from existing benchmarks: (a) *multifaceted* entities (entities that belong to multiple concepts — *e.g.*, 'venice beach' can belong to concepts location and nearby attractions); (b) *non-named* entities (entities that are typically noun phrases but not proper names — *e.g.*, 'coffee'); and (c) *vague* entities (human annotators have subjective disagreement on their concept labels — *e.g.*, 'casino' for nearby attraction). We explain why these characteristics emerge in user-generated text in Section 3.

We found that user-generated text can have up to $10\times$ more *multifaceted* entities and $2\times$ more *non-named* entities compared to well-curated benchmarks. Furthermore, concepts that do not have well-defined semantics result in *vague* entities. We use these characteristics to profile ESE methods, showing that the performance difference between well-curated and user-generated text can partially be attributed to these characteristics.

Contributions. To summarize, our key contributions include: a) identifying and verifying several important new characteristics in user-generated text that are not explored in evaluation of existing ESE methods, b) constructing three new user-generated text benchmarks (we publicly release two[1]), c) proposing new metrics for evaluating ESE methods, d) deriving insights through a cross-domain (user-generated text vs. well-curated) comparison study on different ESE methods.

Key findings. Our main findings are listed below:

- Widely used evaluation metrics such as (mean average precision (MAP) at $k \leq 20$) is an inadequate indicator of the performance of ESE methods on both well-curated and user-generated text. Evaluating top-k_g[2] predictions is potentially more robust, especially for benchmarking.
- Performance of state-of-the-art (SOTA) ESE methods drops dramatically on user-generated text compared to well-curated text.
- Deviating from prior observations, simple corpus-based and language model-based methods that underperform SOTA methods on well-curated text can outperform SOTA methods on user-generated text.
- Simple rank-based ensemble methods can provide further improvements on user-generated text. The degree of overlap of correct predictions from candidate methods is indicative of the effectiveness of their ensemble.

2 Background and Related Work

We now introduce the task of entity set expansion (ESE), existing paradigms and evaluation methods.

2.1 Problem Definition

Given a textual corpus and a user-defined seed set of entities (*e.g.*, 'coffee', 'table') of concepts (*e.g.*, room features), the task of ESE is to output a ranked list of entities (*e.g.*, 'clock', 'tv') that belong to the same concept. Following previous work, we focus on the *low-resource setting* where the seed set is small (3-10 entities per concept).

2.2 Entity Set Expansion Paradigms

To expand the seed set, ESE methods rank candidate entities extracted from a textual corpus (Shang et al., 2018). We limit our scope to low-resource setting and exclude methods (Mao et al., 2020; Takeoka et al., 2021) that require large training examples sub-concepts hierarchy or external knowledge from ontologies and knowledge bases. We organize ESE methods into the following categories.

Corpus-based Methods. These methods (Huang et al., 2020b; Shen et al., 2017, 2018; Yu et al., 2019a) obtain contextual features and distributed representations of entity candidates from the corpus and use them to estimate similarity of candidates to entities in the seed set. This is either done in a single step (Mamou et al., 2018; Yu et al.,

[1] https://github.com/megagonlabs/eseBench

[2] k_g denotes the actual entity set size of a concept.

Characteristics	Examples	Comments
Multifaceted entities	**R1**: Be sure to book in advance an early morning trip to Alcatraz, go to fisherman's wharf . . . **R2**: I would not stay here again. I'd rather pay more and stay by fisherman's wharf . . .	R1 refers to concept *nearby_attraction* while R1 refers to *location*. Entities that fall into multiple semantic concepts might influence other entity candidates for a target concept.
Vague entities	**R1**: . . . see the majestic Frenchy-looking civic center surrounded since 8pm by a crowd . . . **R2**: The Monticello Inn is five to ten minute cab ride from civic center . . .	R1 indicates the entity of interest is a nearby attraction but R2 is vague. Popular concepts such as nearby attractions in user generated text can be inherently subjective.
Non-named entities	**R1**: There was tea and coffee available round the clock in the lounge. **R2**: The room rate included a large and varied continental breakfast with excellent coffee. {concept: room features}	Concepts of interest in user-generated text domain often exhibit non-named entities. In fact, user-generated Tripadvisor dataset has 1.7× more non-named entities (such as "coffee") compared to well-curated Wiki dataset.

Table 1: Exploring different characteristics of well-curated and user-generated text domains.

2019a) or iteratively (Shen et al., 2018; Huang et al., 2020b; Yan et al., 2021).

Language Model-based Methods. Studies have shown that pre-trained language models (LMs) can be used as knowledge bases when queried with prompts (Petroni et al., 2019; Liu et al., 2021). Following this, ESE methods (Zhang et al., 2020a; Takeoka et al., 2021) probe an LM to rank entity candidates. These methods rely on knowledge stored in LMs instead of using them to obtain contextualized representations of entities in the corpus.

Ensemble Methods. CaSE (Yu et al., 2019b) combines context feature selection with pre-trained word embeddings to compute similarities between entities. A similar mechanism, mean reciprocal ranking (MRR) ensemble, has been shown to be effective in combining rankings from different features, views or subsets of seeds (Shen et al., 2017; Zhang et al., 2020b; Huang et al., 2020b).

2.3 Benchmark and Evaluation Metrics

Widely-used benchmarks for ESE, such as Wiki and APR (Shen et al., 2017), are based on well-formed text corpora like Wikipedia and focus only on well-defined concepts such as countries, US states, and diseases. Furthermore, the ranked expansion results are evaluated against the ground truth using Mean Average Precision (MAP) at different top-k positions where k is much smaller than the size of entity set. For example, there are 195 countries but only 10-50 predictions are evaluated. In following sections, we argue that existing benchmarks and evaluation metrics may not be adequate enough to estimate the real-world performance of the ESE methods and introduce new benchmarks and evaluation metrics to address their limitations.

3 Case Study

Existing work suggest that user-generated text differs from well-curated text in writing style (Bražinskas et al., 2020; Huang et al., 2020d) and cleanliness (Van der Wees et al., 2015; Dey et al., 2016). In this section, we discuss one of the use cases of ESE for a downstream NLP application and highlight new characteristics in user-generated text that are particularly relevant to the ESE task.

3.1 Motivating Example

Let us consider a scenario where Tajin, a data scientist at an online travel company (similar to TripAdvisor), has to develop a semantic search feature that helps users explore relevant reviews corresponding to their queries. For example, when a user searches for 'amenities' at a hotel, the feature should display reviews with mentions of different amenities highlighted. Since the reviews are unlabeled, Tajin first consults an expert to compile a list of frequently queried concepts and corresponding example entities that may appear in the reviews (similar to Figuure 1). To discover more entities for each concept, she formulates it as an ESE task, where the goal is to achieve a high coverage of entities in the reviews.

Given a review corpus and seed, Tajin employs a state-of-the-art ESE method that has been evaluated on well-curated text. She finds that it does not perform well in achieving a high coverage of entities in the hotel domain and wonders why. To explain her findings, we explore the characteristics of the TripAdvisor (Miao et al., 2020) and Wiki datasets next and discuss the potential factors that may impact the performance of the SOTA method.

3.2 Observations

Multifaceted entities. Unlike in Wiki benchmark where concepts are well-defined, concepts in Tripadvisor are domain-specific and can have overlapping semantics (see Figure 1). As a result, an entity can belong to multiple concepts. For example, in Table 1, the entity 'fisherman's wharf' can be both location of and nearby attraction to a hotel. We refer to such entities as *multifaceted entities* (Rong et al., 2016). The overlapping semantics of entities can pose challenges to ESE methods that expand multiple concepts simultaneously.

Vague entities. Concept definitions in Wiki benchmark are strict (*e.g.*, countries, states) and ground truth about concept-entity pairs can be obtained by referring to external resources or commonsense. However, some concepts in Tripadvisor are open-ended and subjective, leading to vagueness in interpretation. For example, in Table 1, the terms "nearby" and "attraction" in the concept nearby attractions are subjective. An entity 'civic center' may be neither an attraction nor a nearby one depending on the context in the review. As a result, human annotators independently labeling the entity may disagree on the ground truth label. We refer to entities with subjective disagreements between annotators as *vague entities*. Intuitively, ESE methods may find it difficult to learn to disambiguate the context of vague entities.

Non-named entities. Non-named entities (*e.g.*, 'coffee' and 'tv unit') are typically noun phrases that are not proper names (Paris and Suchanek, 2021). Recent studies (Mbouopda and Melatagia Yonta, 2020; Bamman et al., 2019) have identified that non-named entities are prevalent even in well-curated domains and yet are ignored in existing benchmarks. Non-named entities are even more prevalent in user-generated text. As shown in Table 1, Tripadvisor benchmark contains almost $2\times$ non-named entities than Wiki. Since non-named entities are not canonicalized and can have broader semantics, they can make the ESE task more challenging.

Evaluation Metric. Existing evaluation metrics only consider top 10-50 entities for each target concept (Shen et al., 2017; Zhang et al., 2020a). There are multiple limitations of these metrics. First, there may not be sufficient representation of multi-faceted, vague, and non-named entities in a small set ($<$50 entities). Second, the actual number of correct entities per concept (referred to as *concept size*) may be much larger or smaller than 50. For example, both Tripadvisor and Wiki have larger and varying concept sizes with the median size being 121 and 205, respectively (check Table 2 for more detailed statistics). As a result, focusing only on precision of a small, fixed set of predictions may not reflect the recall of correct entities with respect to concept size.

3.3 Discussion

While multifaceted, vague, and non-named entities can be present in well-curated data, the corresponding benchmarks and downstream applications target real-world named entities and ignore non-named entities (Paris and Suchanek, 2021). In contrast, in most user-generated text domains, the concepts of interest for downstream applications (semantic search feature as discussed above) are not limited to named-entities only and may exhibit multifaceted, vague, and non-named entities (*e.g.*, facts about a hotel such as amenities and attractions). With the increasing use of user-generated text in NLP applications (Xu et al., 2021), it is therefore important to investigate the impact of the aforementioned characteristics on the performance of the ESE methods.

4 Experimental Set-up

We now outline our experiment set-up designed to explore the suitability of existing benchmarks, metrics, and methods.

4.1 Methods

We first describe the ESE methods we evaluate. Following prior work (Shen et al., 2018; Zhang et al., 2020a), we use AutoPhrase (Shang et al., 2018) to generate candidate entity lists from the corpus of a given domain. We then use the following representative publicly available ESE methods from different paradigms to expand the seed set[3].

SetExpan. SetExpan (Shen et al., 2017) is a SOTA corpus-based method that iteratively ranks entity candidates by filtering out noisy skip-gram features. It incorporates other context features such as POS tags and syntactic head tokens in ranking.

[3] All methods were released under Apache 2.0 license.

Embedding baseline (Emb-Base). In order to make use for more robust context embeddings, we develop a simple baseline that uses a pre-trained language model (LM) to derive context embeddings of entity candidates. To derive an entity embedding, we average context embedding of the sentences that mention the entity using BERT (Devlin et al., 2018). We compute concept embeddings by averaging embeddings of its seed entities, and rank entity candidates based on the cosine similarity of concept and entity embeddings.

CGExpan. CGExpan (Zhang et al., 2020b) is a SOTA LM-based method that iteratively uses Hearst patterns (Hearst, 1992) as prompts to obtain scores for ranking candidates. In addition, it considers how a candidate in turn ranks the target concept name to improve the quality of rankings.

LM Probing Baseline (LM-Base). We develop a simpler baseline that also uses Hearst patterns to prompt LMs and obtain scores for entity candidates. However, it does not include any other mechanisms such as concept name guidance and iterative expansion like CGExpan.

Ensemble Methods. We use mean reciprocal rank (MRR) as the representative for ensemble methods since it does not require any additional training data. Given the rankings from multiple methods, we compute MRR score of each entity: $MRR(e) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r_i(e)}$, where n is the number of methods combined, and $r_i(e)$ is the ranking of entity e under method i. We then re-rank all entities based on their MRR score. In this work, we study combinations of two ESE methods leading to 6 ensembles. We study 4 settings that offer interesting combinations across different paradigms:

- MRR-Baseline: Emb-Base + LM-Base.
- MRR-SOTA: SetExpan + CGExpan.
- MRR-Corpus: SetExpan + Emb-Base.
- MRR-LM-Probe: CGExpan + LM-Base.

4.2 Datasets

We use widely adopted well-curated benchmarks: Wiki and APR (Zhang et al., 2020b; Shen et al., 2017). In addition, we create 3 new benchmarks based on user-generated text from Yelp (Huang et al., 2020c), Tripadvisor (Miao et al., 2020) and a proprietary Jobs dataset. All the datasets are in English. We first select concepts for the seed by referring to the features on the corresponding websites, to ensure their relevance for immediate downstream tasks. For example, we select concepts from various facets such as room type, amenities, and distance from attractions that help visitors search hotels on the Tripadvisor website[4]. Table 2 shows selected concepts in the benchmarks.

Data Collection and Annotation. In order to collect ground-truth to construct benchmarks for new domains, we collect top 200 predictions for each concept from each of the ESE methods described in Section 4.1. The first three authors of the paper labeled the predictions, 1 if a concept-entity pair is correct and 0 otherwise. We consider the majority vote as the final label for a concept-entity pair. For entities with rank > 200, we label the corresponding concept-entity pairs to be all negatives based on our preliminary observations that most of them are incorrect. We release the new benchmarks except for the Jobs dataset.

4.3 Metrics

In order to profile the benchmarks, we compute multifacetedness (m) as the fraction of entities in a benchmark that have been assigned to more than one concept. We compute non-named rate (r) as the fraction of non-named entities in the benchmark. We use Spacy[5] to identify named entities in the benchmarks. To avoid bias in estimating vagueness, we hire two additional in-house annotators who are unfamiliar with the concept definitions and entities. They label the ground truth concept-entity pairs — 1 if correct and 0 otherwise[6]. We compute vagueness (κ) in a benchmark using Fleiss' Kappa (Falotico and Quatto, 2015) which measures agreement among the annotators.

Since the benchmarks we constructed are intended to be comprehensive, we propose to estimate mean average precision (MAP) at *gold-k* (k_g) which equals the concept size, *i.e.*, number of entities in the concept. In comparison to smaller and fixed k, evaluation at k_g has several advantages: (a) it can adapt to different concept sizes and (b) it gives an estimate of recall[7] which is crucial to estimate effectiveness in real-world settings with commonly large concept sizes. Intuitively, using

[4] https://tripadvisor.com

[5] https://spacy.io/usage/linguistic-features

[6] Labeling instructions are included in benchmark release.

[7] $P@k_g = R@k_g = F1@k_g$ because the number of predicted positives and true positives both equal to k_g.

Corpus		Seed			Benchmark			
Dataset	**# Docs**	**Example concepts** (# concepts)	**Avg. seed size**	**# entity cands**	**Concept size**	κ	$\mathbf{m}$	$\mathbf{r}$
Wiki	973k	countries, parties, us states, china provinces, companies, tv channels, diseases, sports leagues (8)	9.875	203322	{51, 205, 446}	–†	0.0141	0.4143
APR	1043k	countries, us states, parties (3)	8.333	78870	{89, 202, 301}	–†	0.0000	0.3649
Yelp	757k	restaurant name, restaurant type, seating arrangement, food category, parking, ambience (14)	4.429	23527	{15, 99, 353}	0.0252	0.0369	0.7995
Tripadvisor	18k	location, property type, style, amenities, room features, room type, nearby attractions, staff (8)	6.625	6842	{31, 121, 244}	-0.1252	0.0908	0.7043
Jobs	318k	company, dress code, job position, pay schedule, benefits, payment option (14)	5.143	8028	{36, 100, 316}	-0.1902	0.0837	0.7957

Table 2: Statistics of datasets: no. of documents, example concepts in the seed, avg. no. of entities in the seed, no. of entity candidates and concept size *{min, median, max}* across different concepts. Statistics of benchmarks: multifacetedness (m), non-named rate (r) and vagueness (κ). †: for well-curated datasets, there is no subjective disagreement since the concept-entity pairs are factually verifiable.

k_g would include more instances of multifaceted, vague and non-named entities that would otherwise be ignored in small k. Notice that in certain real-world scenarios, *e.g.*, developing ESE methods for a new domain, estimating k_g may be difficult, thus previous metrics with smaller and fixed k can be useful. However, for other scenarios, especially for evaluation on benchmarks where the goal is to stress test methods, evaluation at k_g is more appropriate.

5 Findings

We next share our findings from analyzing the ESE methods. Note that all the results are obtained from single run of each experiment.

5.1 Appropriateness of Existing Benchmark and Metrics

Q1. Do we need new benchmarks based on user-generated text?

Table 2 compares the characteristics of the various benchmarks using measures described in Section 4.3. As can be seen, user-generated text benchmarks exhibit a higher degree of multifacetedness (m) and non-named rate (r) compared to well-curated Wiki and APR benchmarks. Moreover, poor agreement between annotators ($\kappa < 0$) indicates the presence of vagueness or subjectivity in user-generated text which does not exist in well-curated benchmarks. While all benchmarks exhibit diversity in concept sizes, the diversity is higher in user-generated text than well-curated benchmarks.

Takeaway 1 *User-generated text benchmarks exhibit more multifaceted entities, non-named entities, and vagueness than well-curated benchmarks.*

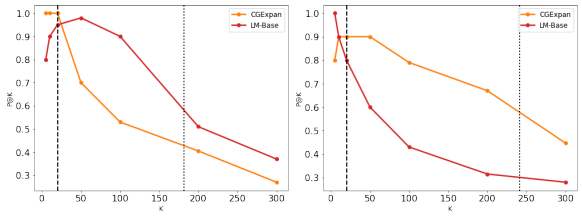
Q2. Do existing evaluation metrics accurately estimate the performance of ESE methods?

Table 3 shows the % drop in MAP of different ESE methods when k is increased from 20 to k_g. The performance drop is consistent across both well-curated and user-generated text benchmarks with the largest being 62% for CGExpan on the Jobs benchmark. This indicates that existing metrics overestimate the real-world performance of all ESE methods. However, simpler baselines, Emb-Base and LM-Base, tend to show lower performance drop than more sophisticated counterparts on user-generated text. This indicates that existing well-curated benchmarks do not reliably capture progress in this field.

Method/Datasets	Jobs	Yelp	TripAdvisor	Wiki	APR
SetExpan	-36.66%	-41.74%	**-43.16%**	-40.76%	-14.25%
CGExpan	**-61.99%**	**-54.37%**	-42.82%	-39.67%	-38.21%
Emb-Base	-54.10%	-43.96%	-36.41%	-35.45%	**-56.33%**
LM-Base	-36.17%	-35.15%	-34.47%	**-56.40%**	-43.96%

Table 3: Drop in performance of different ESE methods from MAP@20 to MAP@k_g. Largest drops in each dataset are highlighted in bold.

We further observed that, across all the benchmarks, the performance drops are higher for concepts with large entity sets. We show two such cases in Figure 2 — one with user-generated text (Figure 2a) and another with well-curated text (Figure 2b) — which illustrate precision curves at different values of k for concepts with large k_g in various benchmarks. As shown, two ESE methods that may show similar performance at k=20 (widely adopted metric) have much larger performance margins at k_g. Thus, evaluation results on only top 20 predictions may be an incomplete depiction of method robustness, especially for concepts with large entity sets.



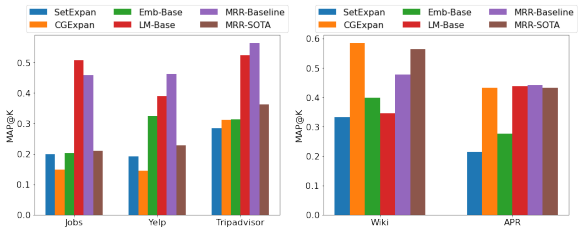
(a) TripAdvisor: amenities (b) Wiki: tv channel

Figure 2: Precision@k of LM-Base and CGExpan for example concepts. Dashed line indicates k=20 and dotted line indicates k_g. Performance margins at k=20 and k_g vary significantly.

Takeaway 2 *Existing evaluation metrics tend to overestimate the real-world performance of ESE methods and may be unreliable for evaluating concepts with large entity sets.*

5.2 Performance on new benchmarks

Q3. How effective SOTA methods are for entity set expansion on user-generated text benchmarks?



(a) user-generated text. (b) well-curated datasets.

Figure 3: Overall MAP@k_g performance of each method on (a) user-generated text and (b) well-curated benchmarks. SOTA methods (CGExpan, SetExpan) are outperformed by simple baselines and ensemble methods on user-generated text.

Given new benchmarks and evaluation metrics, we now compare the performances of various ESE methods. Figure 3 shows that SOTA method CGExpan outperforms other methods on existing well-curated benchmarks which aligns with the reported success of the method. Surprisingly, simpler baseline methods (Emb-Base and LM-Base) that were not optimal on well-curated benchmarks, significantly outperform their SOTA counterparts (SetExpan and CGExpan, respectively) on user-generated text benchmarks, with LM-Base obtaining the best performance. We also observe that ensemble-based methods tend to perform better than or at least similar to the ESE methods they combine.

Takeaway 3 *Performance of SOTA methods do not generalize to user-generated text benchmarks. Ensemble-based methods may improve over the corresponding standalone methods.*

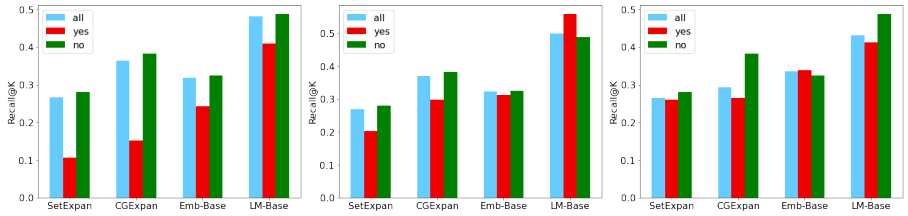
We now examine why SOTA approaches may underperform on user-generated text. Given the success of LM-based contextual representations, it is expected that Emb-Base may outperform lexical feature-based SetExpan. Furthermore, as SetExpan eliminates noisy features of a candidate entity before ranking candidates, it may disregard some context features of multifaceted and vague entities that are mentioned in diverse contexts in user-generated text, leading to sub-optimal ranking of entities. Similarly, CGExpan, which scores each candidate entity by selecting one positive concept and multiple negative concepts, may penalize entities belonging to multiple concepts (multifaceted entities) or mentioned in different contexts (vague entities). Therefore, many of the carefully designed approaches useful on well-curated domains may not generalize to user-generated text.

Takeaway 4 *SOTA methods implement techniques that avoid selecting ambiguous context of an entity. Such a design choice potentially penalizes multifaceted and vague entities when ranking entity candidates for concepts.*

Q4. How do characteristics of user-generated text affect performance of ESE methods?

We now discuss how different characteristics of user-generated text impact the behavior of ESE methods. To understand this, we compare the recall of entities that exhibit one of the target characteristics (multifaceted/non-named/vague) with recall of entities that do not exhibit any of the characteristics. This enables us to analyze the influence of a target characteristic independent of other characteristics. To compute recall, we consider an entity as retrieved if it is ranked in the top-k_g predictions.

Figure 4 compares the recall of entities across different characteristics. For ease of visualization, we combine entities across the 3 benchmarks. As shown, almost all methods show lower recall of entities that exhibit challenging characteristics than entities without these characteristics, and SOTA methods suffer larger drops than simple methods. This supports our hypothesis that characteristics of user-generated text negatively affect performances,



(a) vague entities (b) multifacetedness (c) non-named entities

Figure 4: Recall@k_g of an ESE method typically drops with the presence of (a) vague, (b) multifaceted, and (c) non-named entities in user-generated text (show by red bars) compared to the case when none of these entities are present (shown by green bars).

company	Jobs
Seed: walmart, amazon, subway, microsoft, target	
CGExpan (0.48): costco, apple, AT&T, starbucks, fedex, ALDI, kroger,...	
LMBase (0.45) : nike, apple, IBM, sears, starbucks, target, google, intel,...	

seating arrangement	Yelp
Seed: indoor area, outdoor bar, roof top, patio seating	
CGExpan (0.02): restaurant, atmosphere, live music, music, free wifi, casino..	
LMBase (0.57): bar lounge, outside patio, pool deck, restaurant, rooftop..	

Figure 5: Example concepts on which CGExpan and LM-Base have similar (*company*) or different (*seating arrangement*) performances (indicated in parenthesis). Incorrect entities are shown in red.

especially for SOTA methods which tend to penalize entities with diverse contexts. Future work may investigate how to overcome these challenges.

To provide a qualitative comparison between the behaviors of SOTA methods (*e.g.*, CGExpan) and our proposed baselines (*e.g.*, LM-Base), we show their predictions on two representative concepts in Figure 5. CGExpan and LM-Base have comparable performance on well-formed concepts (*e.g.*, company) in Jobs. However, LM-Base outperforms CGExpan for concepts (*e.g.*, seating arrangement) with entities having characteristics of user-generated text. CGExpan retrieves entities that co-occur frequently with seating arrangement.

Takeaway 5 *Due to the presence of challenging characteristics in user-generated text, performance of all ESE methods are negatively impacted with SOTA methods exhibiting larger drops.*

5.3 Improvement Opportunities

Q5. How do we design ensemble methods for benchmarks with user-generated text?

We analyze ensemble methods further since they tend to outperform other ESE methods (Figure 3). It is trivial that ensemble methods perform well when both combined methods are strong. We are more interested in other factors that may impact performance. Specifically, we investigate what influences the *effectiveness* of a MRR method that combines two ESE methods. An MRR combination is more effective when it outperforms both candidate methods by a larger margin. We define *effectiveness* of combining methods as:

$$Eff(m_1, m_2) = \frac{S(m_1 + m_2)}{\max(S(m_1), S(m_2))} - 1 \quad (1)$$

where $S(m)$ means the performance (MAP@k_g in our study) of method m, and $m_1 + m_2$ means the MRR combination of method m_1, m_2.

As discussed in Section 5.2, multifaceted and vague entities may appear in diverse contexts which SOTA approaches fail to capture, leading to lower recall. Intuitively, it is advantageous to combine methods that capture differing contexts and in the process predict collections of correct entities with minimal overlap. In other words, in order for a MRR method to achieve higher recall, the ESE methods must be *compatible*. We measure *compatibility* of two ESE methods as:

$$Comp(m_1, m_2) = \frac{\|P(m_1) \cup P(m_2)\|}{\max(\|P(m_1)\|, \|P(m_2)\|)} - 1 \quad (2)$$

where $P(m)$ is the set of *correct* entity predictions of method m, i.e. positive benchmark entities ranked among top-k_g by m. $\|\|$ denotes the size of a set. When one of the correct prediction set of m_1, m_2 is a subset of the other, their compatibility is 0. When the two methods find two disjoint sets of correct entities, their compatibility is 1.

We illustrate the correlation between compatibility of method pairs and effectiveness of their MRR combination in Figure 6 using a scatter plot. Each of the points represent the compatibility and effectiveness of the four ensemble methods (MRR-

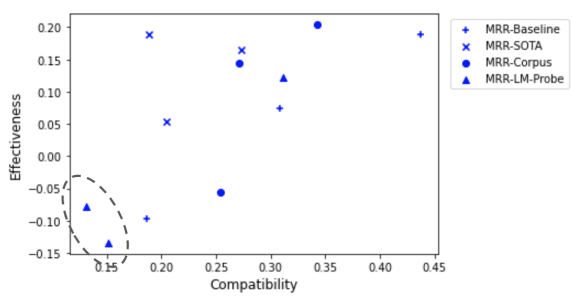


Figure 6: Scatter plot showing the positive correlation between compatibility of methods and effectiveness of their MRR combination on three user-generated text benchmarks. The least compatible and effective combination results (in dashed circle) are from MRR-LM-Probe.

SOTA, MRR-Base, MRR-Corpus, and MRR-LM-probe) on all three user-generated text datasets. We observe that LM-based methods are least compatible due to their similarity in design. The resulting ensemble, MRR-LM-probe, has poor effectiveness (highighted by the dashed ellipse in Figure 6). Other method pairs have less homogeneity in their design and the resulting ensembles often show higher effectiveness. The corresponding compatibility and effectiveness values have a strong positive correlation (Pearson correlation, $R = 0.69$). Therefore, compatibility can be a useful metric for deciding whether combining two methods method may improve performance or not.

Takeaway 6 *Two effective ESE methods on user-generated text with high compatibility (diversity in correct predictions) may achieve higher performance when combined using rank-based ensemble.*

6 Discussion

We now discuss the implications of the proposed benchmark, metrics, and experiment observations.

Capturing the silent majority. Recent work (Paris and Suchanek, 2021) shows that the majority of the entities in Wikipedia articles — which feeds knowledge-bases such as DBpedia (Auer et al., 2007) and YAGO (Suchanek et al., 2007) — are non-named and recommends adding the silent majority to these KBs for completeness. To this end, our proposed benchmark highlights the importance of capturing multifaceted, vague, and non-named entities present in user-generated text. For example, domain-specific KBs such as the Amazon Product Knowledge Graph (Karamanolakis et al., 2020) rely on user-generated text to collect entities for concepts of interest. These KBs power many downstream tasks such as semantic search, question answering, and conversational AI. Therefore, these KBs would remain incomplete without capturing the different types of entities identified in our benchmark.

Practical usage. The goal of our evaluation metrics (evaluation at k_g) is to characterize the performance of ESE methods in the presence of entity types that are typically present in user-generated text. Note that we do not recommend replacing the existing metric MAP@$K = 20$. Our proposed MAP@K_g metric is complementary and is designed to stress test ESE methods in scenarios where coverage is an important criteria (*e.g.*, KB population.) Top-20 predictions do not have enough representation of non-named, multifaceted, and vague entities. Therefore, when evaluating ESE methods designed for user-generated text on our benchmark, the proposed evaluation at k_g metric may help practitioners measure the suitability of a method.

Towards domain-specific ESE. Our study highlights that compared to simple ESE baselines, SOTA methods exhibit poor performance on user-generated text. On the other hand, for well-curated text, SOTA methods outperform the baselines. However, the purpose of this study is not to show that there are better approaches than SOTA methods. Instead, we draw attention to the fact that, there is potential for future research on developing methods for user-generated text domain.

7 Conclusion

We conduct a comprehensive study to analyze the performance of ESE in user-generated text. We observe that user-generated text has characteristics that are not captured in existing benchmarks, and propose new benchmarks and evaluation metrics. Our findings indicate that state-of-the-art methods are not very effective in user-generated text and are often outperformed by simpler baselines.

Acknowledgements

We thank Tom Mitchell and Yoshihiko Suhara for their valuable feedback on the work.

References

Tareq Al-Moslmi, Marc Gallofré Ocaña, Andreas L Opdahl, and Csaba Veres. 2020. Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, 8:32862–32881.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proc. NAACL-HLT' 2019*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikita Bhutani, Aaron Traylor, Chen Chen, Xiaolan Wang, Behzad Golshan, and Wang-Chiew Tan. 2020. Sampo: Unsupervised knowledge base construction for opinions and implications. In *Proc. AKBC' 2020*.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. In *Proc. EMNLP' 2020*.

Ramji Chandrasekaran, Harsh Nilesh Pathak, and Tae Yano. 2020. Deep neural query understanding system at expedia group. In *Proc. IEEE Big Data' 2020*, pages 1476–1484. IEEE.

Pengxiang Cheng and Katrin Erk. 2020. Attending to entities for better text understanding. In *Proc. AAAI' 2020*, 05, pages 7554–7561.

Hongliang Dai and Yangqiu Song. 2019. Neural aspect and opinion term extraction with mined rules as weak supervision. *arXiv preprint arXiv:1907.03750*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2016. A paraphrase and semantic similarity detection system for user generated short-text content on microblogs. In *Proc. COLING' 2016*, pages 2880–2890, Osaka, Japan. The COLING 2016 Organizing Committee.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2):463–470.

Karan Goel, Laurel Orr, Nazneen Fatema Rajani, Jesse Vig, and Christopher Ré. 2021. Goodwill hunting: Analyzing and repurposing off-the-shelf named entity linking systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 205–213, Online. Association for Computational Linguistics.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. COLING' 1992*.

Jiaxin Huang, Yiqing Xie, Yu Meng, Jiaming Shen, Yunyi Zhang, and Jiawei Han. 2020a. Guiding corpus-based set expansion by auxiliary sets generation and co-expansion. In *Proc. WWW' 2020*, pages 2188–2198.

Jiaxin Huang, Yiqing Xie, Yu Meng, Jiaming Shen, Yunyi Zhang, and Jiawei Han. 2020b. Guiding corpus-based set expansion by auxiliary sets generation and co-expansion. *Proc. WWW' 2020*.

Jiaxing Huang, Yiqing Xie, Yu Meng, Yunyi Zhang, and Jiawei Han. 2020c. Corel: Seed-guided topical taxonomy construction by concept learning and relation transferring. *Proc. SIGKDD '2020*.

Rongtao Huang, Bowei Zou, Yu Hong, Wei Zhang, Aiti Aw, and Guodong Zhou. 2020d. Nut-rc: Noisy user-generated text-oriented reading comprehension. In *Proc. COLING' 2020*, pages 2687–2698.

Emrah Inan, Paul Thompson, Tim Yates, and Sophia Ananiadou. 2021. Hsearch: Semantic search system for workplace accident reports. In *Proc. ECIR 2021*, volume 12657 of *Lecture Notes in Computer Science*, pages 514–519. Springer.

Giannis Karamanolakis, Jun Ma, and Xin Luna Dong. 2020. Txtract: Taxonomy-aware knowledge extraction for thousands of product categories. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8489–8502.

Fatemeh Lashkari, Ebrahim Bagheri, and Ali A Ghorbani. 2019. Neural embedding-based indices for semantic search. *Information Processing & Management*, 56(3):733–755.

Yuliang Li, Aaron Xixuan Feng, Jinfeng Li, Saran Mumick, Alon Halevy, Vivian Li, and Wang-Chiew Tan. 2019. Subjective databases. *arXiv preprint arXiv:1902.09661*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Jonathan Mamou, Oren Pereg, Moshe Wasserblat, Alon Eirew, Yael Green, Shira Guskin, Peter Izsak, and Daniel Korat. 2018. Term set expansion based nlp architect by intel ai lab. *arXiv preprint arXiv:1808.08953*.

Yuning Mao, Tong Zhao, Andrey Kan, Chenwei Zhang, Xin Luna Dong, Christos Faloutsos, and Jiawei Han. 2020. Octet: Online catalog taxonomy enrichment with self-supervision. In *Proc. SIGKDD' 2020*, pages 2247–2257.

Michael Franklin Mbouopda and Paulin Melatagia Yonta. 2020. Named entity recognition in low-resource languages using cross-lingual distributional word representation. *Revue Africaine de la Recherche en Informatique et Mathematiques Appliquees*, 33.

Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan. 2020. Snippext: Semi-supervised opinion mining with augmented data. In *Proc. WWW' 2020*, pages 617–628.

Pierre-Henri Paris and Fabian M Suchanek. 2021. Non-named entities-the silent majority. In *Proc. ESWC' 2021 Poster and Demo Track*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Xin Rong, Zhe Chen, Qiaozhu Mei, and Eytan Adar. 2016. Egoset: Exploiting word ego-networks and user-generated ontology for multifaceted set expansion. *Proc. WSDM' 2016*.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *Proc. TKDE' 2018*, 30:1825–1837.

Jiaming Shen, Zeqiu Wu, Dongming Lei, Jingbo Shang, Xiang Ren, and Jiawei Han. 2017. Setexpan: Corpus-based set expansion via context feature selection and rank ensemble. In *Proc. ECML PKDD' 2017*, pages 288–304. Springer.

Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T Vanni, Brian M Sadler, and Jiawei Han. 2018. Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion. In *Proc. SIGKDD' 2018*, pages 2180–2189.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Kunihiro Takeoka, Kosuke Akimoto, and Masafumi Oyamada. 2021. Low-resource taxonomy enrichment with pretrained language models. In *Proc. EMNLP' 2021*, pages 2747–2758.

Marlies Van der Wees, Arianna Bisazza, and Christof Monz. 2015. Five shades of noise: Analyzing machine translation errors in user-generated text. In *Proc. Workshop on Noisy User-generated Text' 2015*, pages 28–37.

Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors. 2021. *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*. Association for Computational Linguistics, Online.

Lingyong Yan, Xianpei Han, and Le Sun. 2021. Progressive adversarial learning for bootstrapping: A case study on entity set expansion. In *Proc. EMNLP' 2021*, pages 9673–9682, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Puxuan Yu, Zhiqi Huang, Razieh Rahimi, and James Allan. 2019a. Corpus-based set expansion with lexical features and distributed representations. In *Proc. SIGIR' 2019*, pages 1153–1156.

Puxuan Yu, Zhiqi Huang, Razieh Rahimi, and James Allan. 2019b. Corpus-based set expansion with lexical features and distributed representations. *Proc. SIGIR' 2019*.

Yunyi Zhang, Jiaming Shen, Jingbo Shang, and Jiawei Han. 2020a. Empower entity set expansion via language model probing. *arXiv preprint arXiv:2004.13897*.

Yunyi Zhang, Jiaming Shen, Jingbo Shang, and Jiawei Han. 2020b. Empower entity set expansion via language model probing. *ArXiv*, abs/2004.13897.