FUSE: MULTI-FACETED SET EXPANSION BY COHERENT CLUSTERING OF SKIP-GRAMS

BY

WANZHENG ZHU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Adviser:

Professor Jiawei Han

ABSTRACT

Set expansion aims to expand a small set of seed entities into a complete set of relevant entities. Most existing approaches assume the input seed set is unambiguous and completely ignore the multi-faceted semantics of seed entities. As a result, given the seed set {"Canon", "Sony", "Nikon"}, previous methods return *one mixed set* of entities that are either camera brands or Japanese companies. In this thesis, we study the task of **multi-faceted set expansion**, which aims to capture all semantic facets in the seed set and returns *multiple sets* of entities, one for each semantic facet. We propose an unsupervised framework, FUSE, which consists of three major components: (1) *facet discovery module*: identifies all semantic facets of each seed entity by extracting and clustering its skip-grams, (2) *facet fusion module*: discovers shared semantic facets of the entire seed set by an optimization formulation, and (3) *entity expansion module*: expands each semantic facet by utilizing an iterative algorithm robust to skip-gram noise. Extensive experiments demonstrate that our algorithm, FUSE, can accurately identify multiple semantic facets of the seed set and generate quality entities for each facet.

To my parents, for their love and support.

ACKNOWLEDGMENTS

I would like to thank my advisor Professor Jiawei Han of the Department of Computer Science at University of Illinois at Urbana-Champaign. He has been helping me with insightful discussions, inspiring brainstorming and immense knowledge. He constantly encouraged me to brainstorm new problems, provided me valuable suggestions and supported me along my entire research path.

I would also like to thank Professor Chao Zhang, for his inspiring discussions as well as his help on conducting experimental studies. Also, I would like to thank Jiaming Shen, Jingbo Shang, Hongyu Gong, Yu Zhang and all my fellow members in the Data Mining Group at the University of Illinois at Urbana-Champaign. I really have learned a lot from you and enjoyed my time with you.

Finally, I would like to thank my parents for their love and support throughout my life.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The task of *set expansion* is to expand a small set of seed entities into a more complete set of relevant entities. For example, to explore all *Universities in the U.S.*, one can feed a seed set (*e.g.*, {"Stanford", "UCB", "Harvard"}) to a set expansion system and then expect outputs such as "Princeton", "MIT" and "UW". Those expanded entities can benefit numerous entity-aware applications, including query suggestion [1], taxonomy construction [2], and information extraction [3, 4]. Besides, the set expansion algorithm itself becomes a basic building block of many natural language processing based systems [5].

Previous studies on set expansion focus on returning *one single set* of most relevant entities. Methods have been developed to incrementally and iteratively add the entities of high confidence scores into the set. A variety of features are extracted, including word co-occurrence statistics [6], unary patterns [7], or coordinational patterns [8], from different data sources such as query log [9], web table [10], and raw text corpus [7, 5]. However, all these methods assume the given seed set is unambiguous and completely ignore the multi-faceted semantics of seed entities. As a result, given a seed set {"apollo", "artemis", "poseidon"} which has two semantic facets – *Major gods in Greek mythology* and *NASA missions*, previous methods can only generate one mixed set of entities from these two facets, which inevitably hampers their applicabilities.

In this thesis, we approach the set expansion task from a new angle. Our study focuses on **multi-faceted set expansion** which aims to identify semantic facets shared by all seed entities and return multiple expanded sets, one for each semantic facet. The key challenge lies in the discovery of shared semantic facets from a seed set. However, the only initial attempt toward multi-facetedness, EgoSet [11], requires user-created ontologies as external knowledge and does not guarantee that their generated semantic facets are

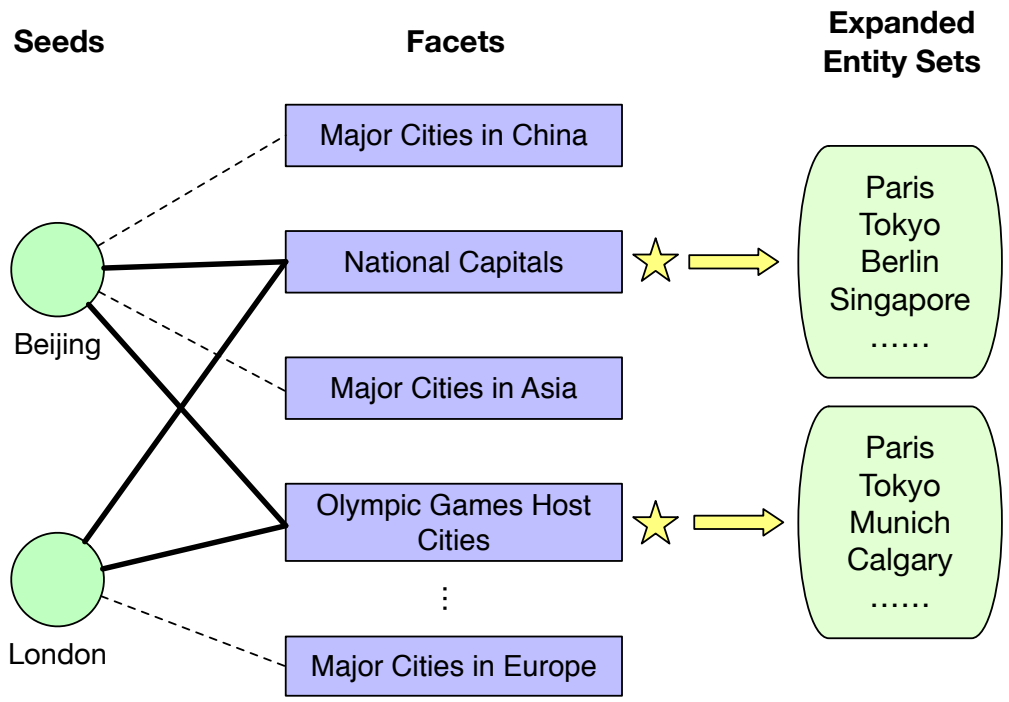

Figure 1.1: An illustrative example of a multi-faceted seed set {"Beijing", "London"}. Facets (*e.g. Major Cities in China*) that do not appear in both seed entities should be eliminated in set expansion. As a result, we expect to output two separated entity sets: one for facet *National Capitals* and the other one for facet *Olympic Games Host Cities*.

relevant to all seed entities. As an example in Fig. 1.1, EgoSet generates more than five facets, but only two of them are relevant to both seeds.

To handle the key challenge of multi-faceted set expansion, we propose a novel framework, FUSE, as illustrated in Fig. 1.2. First, we discover all possible facets of each seed by extracting and clustering its skip-grams. Second, we leverage an optimization formulation to discover the shared semantic facets across all seeds as *coherent semantic facets*. This helps eliminate those facets relevant only to a partial set of seeds. Third, based on the coherent skip-gram clusters, we design an iterative framework to further reduce skip-gram noise and provide quality entities for each facet.

It is considerably complicated to evaluate such multi-faceted set expansion task, mainly because we have no prior knowledge about the number of facets in a seed set. Therefore, we are likely to observe a different number of facets between the generated result and the ground truth (*e.g.*, the ground truth may have three facets, while the generated result has four facets).

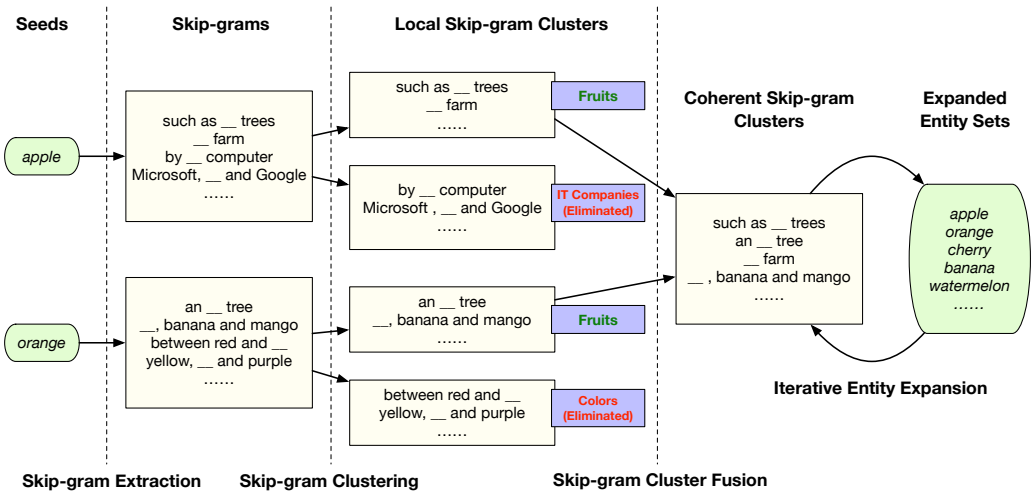

Figure 1.2: Overview of FUSE. The key novelty is to discover coherent skip-gram clusters, whereas previous methods skip this stage and directly combine all skip-grams into the same pool for the expansion.

Previously proposed metric MMAP in [11] only measures how many entities and facets in the ground truth are covered by the generated result. However, it fails to measure how noisy those generated facets are and thus it biases toward methods that output as many facets as possible. To overcome the intrinsic limitation of MMAP, we propose a more comprehensive evaluation metric, BMAP, that can capture both the purity of generated facets and their coverage of ground truth facets.

Our contributions are highlighted as follows.

- We identify the key challenge of a new problem – *multi-faceted set expansion* and develop an iterative framework, FUSE, to address it.[1]
- We propose a new evaluation metric for the multi-faceted set expansion problem, which is shown to be a more comprehensive measure.
- Extensive experiments demonstrate that our proposed framework outperforms state-of-the-art set expansion algorithms significantly in both accuracy and efficiency.

[1] `https://github.com/WanzhengZhu/SetExpan-MultiFacet`

CHAPTER 2

PROBLEM FORMULATION

A facet refers to one semantic aspect or sense of seed words. For example, *fruit* and *company* are two facets of the word "apple". Previous works study mostly single-faceted set expansion and ignore the seeds' multi-facetedness nature. In this work, we explore a better coverage of all coherent semantic facets of a seed set and study corpus-based multi-faceted set expansion.

More formally, given a seed set query $q = \{s_1, s_2, ..., s_m\}$ where s_i is a seed and a raw text corpus D, our set expansion system is to find all lists of entities $\mathbb{E} = \{E^{(i)}, E^{(j)}, E^{(k)}, \dots\}$, where $E^{(i)} = \{x_1^{(i)}, \dots, x_n^{(i)}\}$ is relevant to the i-th facet f_i of query q, and $x_l^{(i)}$ denotes an expanded entity.

CHAPTER 3

MODEL

Our proposed FUSE framework consists of three main steps: (1) extracting and clustering skip-grams for each seed (c.f. Section 3.1); (2) discovering coherent semantic facets of a seed set (c.f. Section 3.2); and (3) iteratively expanding entity sets for each semantic facet (c.f. Section 3.3).

An overview of our approach is shown in Fig. 1.2 and the algorithm is shown in Algorithm 1.

3.1 Skip-Gram Features Extraction and Clustering

We preprocess the raw corpus as SetExpan [7] does, and extract skip-gram features of seed words as Egoset [11] does. Here skip-gram features are a sequence of words surrounding the seed word. Based on the distributional hypothesis [12], the semantics of the seed word is reflected by its neighboring skip-grams. We can derive different facets of a seed word by separating its skip-grams into different semantic clusters.

Embedding is commonly used in natural language processing applications to represent rich semantic information of words and phrases. We obtain the embedding for each skip-gram by simply averaging the embedding of its component words. The derivation of skip-gram embedding is another interesting research question, but it is not our focus in this work.

Now we cluster these skip-gram embeddings to discover different semantic facets of a seed word. Most clustering algorithms require the number of clusters as input, which deviates from our problem setting. Also, we note that the embedding usually lies in a high-dimension space (typically of dimension 100-300), which leads to the poor and unstable performance of existing non-parametric clustering algorithms [13] (*e.g.*, MeanShift [14]).

One of our contributions is to cast the skip-gram clustering problem as a

graph community detection problem. By doing this, we can get rid of the two main issues of instability and hard coded cluster numbers as mentioned above. Specifically, we will construct a complete weighted graph where each node represents a skip-gram, and the edge weight between each pair of nodes indicates the cosine similarity of their corresponding skip-gram embeddings. As the edge weight measures the semantic relevance of two skip-grams, we expect those semantically coherent skip-grams are put into one graph cluster (*i.e.*, a community in graph). Given a graph cluster C with skip-gram clusters C_i, C_j, ..., we define its intra-cluster relevance $R(C)$ as the total relevance between word pairs in the same cluster as follows:

$$R(C) = \sum_{i \neq j} r_{i,j} \cdot \mathbf{1}_{(C_i = C_j)} \tag{3.1}$$

where $r_{i,j}$ is the relevance between skip-gram i and j (defined in Eq. 3.2), and C_i, C_j are the clusters of node i and j, respectively.

Translating this goal into graph model, we want to group nodes with strong semantic connections to each other. We do not use the edge weight as connection strength directly, since some nodes might have high edge weight with all nodes while other nodes tend to have low edge weights. We use a normalized edge weight as the relevance of the node pair (i, j) by subtracting node-specific weight bias from the edge weight $e_{i,j}$. Mathematically, we define the node relevance as follows:

$$r(i, j) = e_{i,j} - \frac{1}{W} \cdot (\sum_k e_{i,k}) \cdot (\sum_k e_{j,k}) \tag{3.2}$$

where i, j are two nodes, $e_{i,j}$ is the edge weight, $\sum_k e_{i,k}$ is the total edge weights of node i, and W is the total edge weight in the graph. Finally, by replacing $r(i, j)$ in Eq. 3.1 with Eq. 3.2, we can obtain the intra-cluster relevance $R(C)$ of the given cluster C.

To find cluster C that maximizes the intra-cluster relevance $R(C)$, we adopt the graph community detection algorithm Louvain [15]. Louvain starts by assigning a different community to each node. Then, it greedily aggregates communities to optimize the intra-community relevance until the relevance cannot be further improved, by when the "optimal" number of communities

will be identified. Empirical results justify that the community detection based skip-gram clustering is able to identify a reasonable number of facets (c.f. Section 4.4), without setting any parameter or threshold.

3.2 Discovering Coherent Semantic Facets of a Seed Set

After obtaining multiple skip-gram clusters for each seed, we then need to find the shared semantic facets among all seeds and generate the coherent skip-gram clusters. Take two seed words "apple" with facets *fruit* and *company*, and "orange" with facets *fruit* and *color* as an example, their common facet is *fruit*.

The key is to determine whether a facet of seed word A matches any facet of word B. Suppose that A has r skip-gram clusters $S_A = \{S_A^{(1)}, \ldots, S_A^{(r)}\}$ where cluster $S_A^{(i)}$ contains a set of skip-grams relevant to the i-th facet of A. Similarly, B has t skip-gram clusters $S_B = \{S_B^{(1)}, \ldots, S_B^{(t)}\}$. If A and B share k facets, and accordingly they have k pairs of matching clusters $\{(S_A^{(i_1)}, S_B^{(j_1)}), \ldots, (S_A^{(i_k)}, S_B^{(j_k)})\}$. Therefore, these k facets are jointly represented by these clusters: $S_{A,B} = \{S_A^{(i_1)} \bigcup S_B^{(j_1)}, \ldots, S_A^{(i_k)} \bigcup S_B^{(j_k)}\}$.

We first measure the pairwise correlation of their skip-gram clusters (c.f. Section 3.2.1), and then make a matching decision on a pair of clusters (c.f. Section 3.2.2).

3.2.1 Calculating correlation between two skip-gram clusters

Suppose that facet A_1 (one facet of word A) corresponds to a skip-gram cluster $\mathbf{X} = [\mathbf{x}_1; \ldots; \mathbf{x}_m]$ with m skip-gram vectors, where $\mathbf{x}_i \in \mathbb{R}^d$. Similarly, facet B_1 (one facet of word B) corresponds to a cluster $\mathbf{Y} = [\mathbf{y}_1; \ldots; \mathbf{y}_n]$ with n skip-gram vectors, where $\mathbf{y}_j \in \mathbb{R}^d$. Two clusters $\mathbf{X}$ and $\mathbf{Y}$ are from different seed words, and we want to measure their correlation in order to decide whether they correspond to the same semantic facet.

To measure their correlation, we find the semantic sense which $\mathbf{X}$ and $\mathbf{Y}$ have in common. Inspired by the idea of compositional semantics [16, 17], we set the sense vector to the linear combination of skip-gram vectors.

Suppose that the sense vector $\mathbf{u}$ from cluster $\mathbf{X}$ and the sense vector $\mathbf{v}$ from $\mathbf{Y}$ are the sense shared by the two clusters. Therefore, the common sense vectors should be highly correlated, *i.e.*, we want to find $\mathbf{u}$ and $\mathbf{v}$ so that their correlation $\mathbf{u}^T\mathbf{v}$ is maximized. We formulate the following optimization problem in Eq. 3.3.

$$\begin{aligned} & \max_{a,b} \frac{\mathbf{u}^T\mathbf{v}}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|} \\ \text{s.t.} \quad & \mathbf{u} = \mathbf{X}\mathbf{a}, \\ & \mathbf{v} = \mathbf{Y}\mathbf{b}, \end{aligned} \tag{3.3}$$

where $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b} \in \mathbb{R}^n$ are coeffcient vectors.

Solving the problem (3.3) by CCA [18], we can find their common sense vectors $\mathbf{u}^*$ and $\mathbf{v}^*$. The semantic correlation $corr(\mathbf{X}, \mathbf{Y})$ between cluster $\mathbf{X}$ and $\mathbf{Y}$ is defined as the correlation between these two sense vectors:

$$corr(\mathbf{X}, \mathbf{Y}) = \mathbf{u}^{*T}\mathbf{v}^* \tag{3.4}$$

3.2.2 Matching facets of all seeds

After quantifying correlation for two skip-gram clusters, we cast it as a binary decision whether the cluster $\mathbf{X}$ of facet A_1 matches semantically with any facet of word B.

We note that it is not a good way to decide the matching clusters by setting a hard correlation threshold, since the the numerical correlation range is word-specific.

It is easy to see that if a facet of seed A (*e.g.* A_1) is of the same semantic class with a facet of seed B (*e.g.* B_2), then $corr(A_1, B_2)$ is higher than the correlation between A_1 and any other facets of seed B. Otherwise, the correlation of A_1 and all facets of seed B should be equally small.

Based on the intuition above, we define a relevance score as follows:

$$rele(A_1, B) = D_{KL}(\mathbf{Corr}(A_1, B), U) \tag{3.5}$$

where U is uniform distribution, D_{KL} is KL-divergence [19], and $\mathbf{Corr}(A_1, B) = softmax((corr(A_1, B_1), ..., corr(A_1, B_m)))$.

We then make the matching decision based on the relevance score $rele(A_1, B)$.[1] Once the matching decision is satisfied, we find the best matching facet B^* in word B and generate the coherent skip-gram cluster $A_1 \bigcup B^*$.

Remarks: If there are more than two seed words, we first discover coherent skip-gram clusters of two seeds and then use their coherent skip-gram clusters to match with the third seed and so on.

3.3 Iterative Entity Expansion

The skip-gram clusters for different facets are used to expand the seed set by adding their neighboring words as relevant candidates. This is again based on the distributional hypothesis that words co-occurring with similar skip-grams are likely to be semantically similar.

It is unavoidable to include noises in the generated skip-gram clusters due to various factors such as noisy skip-gram representations and the clustering process. We note that noisy skip-grams can seriously degrade the quality of generated words in the set expansion process. In this part, we will illustrate how we can reduce the skip-gram noise with an iterative refining process.

For each skip-gram denoted as sg, we learn its "TF-IDF" weight [11, 7] $h_{c,sg}$ associated with a word candidate c,

$$h_{c,sg} = log(1 + N_{c,sg})[\log \frac{|W|}{\sum_{c'} N_{c',sg}}]$$

where $N_{c,sg}$ is the co-occurrences of word c and skip-gram sg in the corpus, and $|W|$ is the total number of candidate entities.

Therefore, given a set of skip-grams, the importance weight w_c of a word candidate c is:

$$w_c = \sum_{sg'} h_{c,sg'} \cdot w_{sg'}$$

where $w_{sg'}$ is the skip-gram weight defined as:

$$w_{sg} = \sum_{c'} h_{c',sg} \cdot w_{c'}$$

[1] Empirically, we find the results are not sensetive to the threshold and set the relevance score threshold to 0.25 in our experiment.

Algorithm 1: FUSE: Multi-faceted Set Expansion

Input: Corpus D; a user query q.
Output: a list of expanded entity lists $\mathbb{E}$.

⊡ ***Skip-gram Clustering***;
seedClusterDict = {};
for *seed in q* **do**
    sgs ← extractSkipgrams(*seed*, D);
    sgClusters ← clustering(sgs);
    seedClusterDict[*seed*] ← sgClusters;
⊡ ***Clusters Fusing***;
refSeed ← q.pop();
refC ← seedClusterDict[refSeed];
while *seed is not empty* **do**
    curSeed ← q.pop();
    curC ← seedClusterDict[curSeed];
    coherentC ← fuseClusters(refC, curC);
    refC ← coherentC;
⊡ ***Entity Expansion***;
$\mathbb{E}$ ← entityExpansion(refC);
return $\mathbb{E}$;

Two equations above, in fact, represent an iterative framework to update word weights and skip-gram weights respectively. Quality skip-grams will rank quality words high and in turn, quality words will make quality skip-grams weigh more. For the first iteration, all skip-gram weights w_{sg} are set to one. Then we proceed with word weights update and skip-gram weights update, and then iterate. Empirically, we find that it converges very fast and we use three iterations in our experiment.

CHAPTER 4

EXPERIMENTS

Our model targets the corpus-based entity set expansion problem, and thus we evaluate its performance on a local corpus.

Dataset: We evaluate our approach, FUSE, on the public benchmark dataset used in [11]. This dataset contains 56 million articles (1.2 billion words) retrieved from English Wikipedia 2014 Dump and 150 human-labeled multi-faceted queries.

4.1 Evaluation Metric

It is considerably complicated to properly evaluate multi-faceted set expansion task due to different number of facets between the generated result and the ground truth. Previous work [11] adopted the following mean of mean average precision (MMAP) measure:

$$MMAP@l = \frac{1}{M_q} \sum_{m=1}^{M_q} AP_l(B_{qi^*}, G_{qm})$$

where M_q is the number of facets for query q in the ground truth; G_{qm} is the ground truth set of m-th facet for q; B_{qi^*} is the output facet that best matches G_{qm}, and $AP_l(c, r)$ represents the average precision of top l entities in a ranked list c given an unordered ground truth set r. This metric measures the coverage of ground truth sets by the generated sets.

However, it does not penalize additional noisy facets in generated sets and thus it is biased towards the model that generates more facets. For example, a model generating 15 facets with three relevant facets achieves higher MMAP than another model generating three facets with two relevant facets. One can "cheat" the performance by generating as many facets as possible.

To overcome the intrinsic limitation of MMAP, we, inspired by [20, 21],

propose a new metric, Best-Matching Average Precision (BMAP) to capture both the purity of generated facets and their coverage of ground truth facets. Our metric is defined as follows:

$$BMAP@l = HMean(MMAP@l, PMAP@l)$$

$$PMAP@l = \frac{1}{F_q} \sum_{f=1}^{F_q} AP_l(B_{qf}; G_{qi^*})$$

where F_q is the number of facets in generated output; B_{qf} is the f-th output ranked list for query q; G_{qi^*} is the ground truth facet that best matches B_{qf}. Here $HMean(a, b) = \frac{2ab}{a+b}$ is the harmonic mean of a and b.

Our proposed BMAP metric not only evaluates how well generated facets match the ground truth by $MMAP@l$ but also penalizes low-quality facets by $PMAP@l$. Intuitively, $MMAP@l$ measures "recall" to capture how many ground truth results has been discovered, while $PMAP@l$ measures "precision" to capture the fraction of good facets in the generated output. Accordingly, $BMAP@l$ measures "F1 score" to leverage "precision" and "recall". Results are reported by averaging all 150 queries.

4.2 Methods

The following approaches are compared:

word2vec[1] [22]: We use the "skip-gram" model in word2vec to learn the embedding vector for each entity, and then return k nearest neighbors of the seed words.

SEISA [23]: An entity set expansion algorithm based on iterative similarity aggregation. It uses the occurrence of entities in web list and query log as entity features. In our experiments, we replace the web list and query log with skip-gram features.

SetExpan[2] [7]: A corpus-based set expansion that selects quality context features for entity-entity similarity calculation and expand the entity sets using rank ensemble.

EgoSet [11]: The only existing work for multi-faceted set expansion. It ex-

[1] `https://code.google.com/p/word2vec`

[2] `https://github.com/mickeystroller/SetExpan`

pands word entities from skip-gram features, and then clusters the expanded entities into multiple sets.

Sensegram[3] [24]: We learn different embeddings for each word's different senses and return k nearest neighbors for each embedding.

FUSE-k-means: A variant of FUSE which replaces Louvain with k-means clustering algorithm for skip-gram clustering. We set the number of clusters k equals to two and three, which are the mode and the mean of the number of clusters of the ground truth respectively.

FUSE-NoIter: A variant of FUSE without the iterative skip-gram selection module.

4.3 End-to-End Evaluation

We compare the end-to-end performance of FUSE against all baselines using MMAP ("recall"), PMAP ("precision") and BMAP ("F1 score"), shown in Table 4.1. FUSE achieves the highest scores in all cases and outperforms all other baselines with obvious margins in BMAP.

It is worth mentioning that EgoSet achieves decent results in MMAP. However, it generates too many noisy facets, which deteriorate PMAP and the overall performance BMAP. We will further discuss this phenomenon in Section 4.4.

It is also interesting to note that single-faceted baselines (*i.e.*, SetExpan) have much stronger PMAP performance than multi-faceted baselines. This is because by generating a single cluster of the most confident expansion results, they usually match with one ground truth cluster very well and thus achieve high PMAP ("precision") value.

In the ablation analysis, it is worth noting that FUSE, even without predetermined number of clusters, performs better than FUSE-k-means. We think it is because the noise of forcing skip-grams into a fixed number of clusters will propagate to the skip-gram cluster fusing step, and thus lead to bad performance. Meanwhile, the comparison between FUSE and FUSE-NoIter demonstrates that our iterative framework is able to reduce the skip-gram noise and generate more coherent clusters.

[3] `https://github.com/uhh-lt/sensegram`

Table 4.1: End-to-end evaluation.

(a) End-to-end evaluation using MMAP ("recall").

	MMAP@*l*		
	l=5	**l=10**	**l=20**
word2vec	0.323	0.283	0.252
SEISA	0.345	0.301	0.268
SetExpan	0.373	0.337	0.304
Sensegram	0.312	0.301	0.275
EgoSet	0.446	0.390	0.325
FUSE-k-means (k=2)	0.419	0.365	0.328
FUSE-k-means (k=3)	0.444	0.387	0.350
FUSE-NoIter	0.437	0.371	0.333
FUSE	**0.449**	**0.398**	**0.361**

(b) End-to-end evaluation using PMAP ("precision").

	PMAP@*l*		
	l=5	**l=10**	**l=20**
word2vec	0.552	0.499	0.448
SEISA	0.550	0.503	0.455
SetExpan	0.605	0.563	0.512
Sensegram	0.479	0.443	0.398
EgoSet	0.306	0.261	0.206
FUSE-k-means (k=2)	0.607	0.546	0.506
FUSE-k-means (k=3)	0.620	0.550	0.496
FUSE-NoIter	0.601	0.533	0.478
FUSE	**0.643**	**0.570**	**0.517**

(c) End-to-end evaluation using BMAP ("F1 score").

	BMAP@*l*		
	l=5	**l=10**	**l=20**
word2vec	0.390	0.352	0.316
SEISA	0.408	0.368	0.331
SetExpan	0.448	0.413	0.374
Sensegram	0.359	0.343	0.314
EgoSet	0.335	0.292	0.236
FUSE-k-means (k=2)	0.477	0.423	0.388
FUSE-k-means (k=3)	0.500	0.442	0.400
FUSE-NoIter	0.490	0.425	0.382
FUSE	**0.513**	**0.450**	**0.413**

4.4 Number of Facets Identified

We explore the number of facets identified by different multi-faceted set expansion methods. Specifically, we adopt l_1 and l_2 distances.

$$l_1 \text{ distance} = \sum_{q \in Q} |\text{GT}_q - \text{Gen}_q|$$
$$l_2 \text{ distance} = \sqrt{\sum_{q \in Q} (\text{GT}_q - \text{Gen}_q)^2}$$

Here Q is all queries, GT_q and Gen_q are the number of facets that ground truth has and the number of facets that the corresponding model identifies for query q, respectively.

Table 4.2: Distance between number of facets identified and number of facets the ground truth has.

	l_1 distance	l_2 distance
EgoSet	783	78.02
FUSE	**159**	**26.55**

As shown in Table 4.2, FUSE is able to generate closer number of clusters to the ground truth, compared to EgoSet, demonstrating about 80% reduction of the l_1 distance.

4.5 Efficiency Comparison

Computational analysis is shown in Fig. 4.1. Word2vec and Sensegram have the best computational efficiency since they only need to find top-k nearest neighbors. FUSE has the next best computational performance, and largely outperforms all other baselines. The heavy computations of EgoSet, SetExpan and SEISA are the entity-entity similarity calculation for ego-network construction, context feature selection, and iterative similarity aggregation respectively. While in FUSE, we adopt cosine distance of skip-gram embeddings for similarity measurement, and linear combinations of skip-gram-to-entity weights for entity expansion. Both steps in FUSE are very efficient.

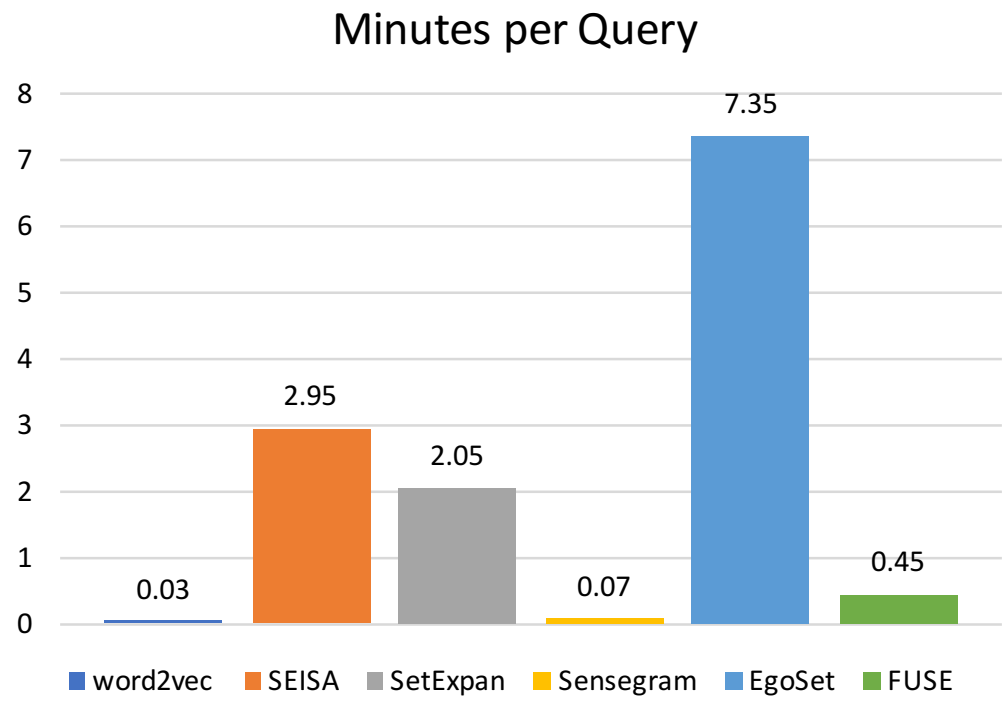

Figure 4.1: Computational efficiency.

4.6 Case Studies

Table 4.3 presents intermediate results of FUSE by showing top skip-grams of each facet. Besides, we present two cases where ground truth results are incomprehensive (Cases 4[4] and 5[5]). It shows that FUSE is able to not only select decent skip-gram features, but also generate quality semantic facets which the ground truth does not even capture.

[4]Beaver River, Elk River and Bear River are tributaries of Pennsylvania, Mississippi River and the Great Salt Lake, respectively.

[5]Chongqing was the second capital of Chinese nationalist party during the war.

Table 4.3: Case study on top skip-grams for each semantic facet. The concept name of each facet is in bold.

ID	Query	Identified Facets: Associated Top Skip-grams	Ground Truth Facets: Example Entities
1	{hydrogen, uranium}	**Chemical Elements**: (helium , __ ,), (of __ and nitrogen.) **Energy Sources:** (for __ energy), (in __ energy), (the __ fuel cell)	**Chemical Elements**: Helium, Carbon, Nitrogen, Oxygen **Energy Sources:** Solar, Coal, Oil, Natural Gas
2	{apollo, artemis, poseidon}	**Major gods**: (the god __ and), (zeus , __ ,), (, athena , __ ,) **NASA missions**: (nasa 's __ ,)	**Major Greek gods**: Aphrodite, Ares, Athena, Zeus **NASA missions**: Juno, Voyager, InSight, NuSTAR
3	{Beijing}	**Chinese Major Cities**: (in __ , china), (, __ , shanghai) **Capitals/International Major Cities**: (paris and __ .), (__ capital) **International Metropolitan/Art Cities**: (theater in __ .), (of music in __ ,), (with the __ symphony orchestra) **Olympic Games Host Cities**: (olympic games in __ ,), (at the __ olympic)	**Chinese Major Cities**: Beijing, Shanghai, Wuhan, Harbin **Province-level divisions of China**: Beijing, Jiangsu, Zhejiang **Capital cities in the world**: Paris, Tokyo, Jakarta, Berlin **Olympic Games Host Cities**: Paris, Tokyo, Munich, Calcary
4	{beaver, elk, bear}	**Animals**: (tailed deer , __ ,), (wolf , __ and) **Tributaries**: (in __ river,), (along the __ river.)	**Animals**: alligator, bear, deer, pig
5	{Chongqing}	**Chinese Major Cities**: (of __ , china.), (based in __ , china.) **War related Major Cities**: (__ broadcasting), (congress of __ .), (party in __ led by)	**Chinese Major Cities**: Beijing, Shanghai, Wuhan, Harbin **Province-level divisions of China**: Beijing, Jiangsu, Zhejiang, Guangdong

CHAPTER 5

RELATED WORK

Early work on entity set expansion, including *Google Sets* [9] and *SEAL* [25] submits a query consisting of seed entities to a general-domain search engine (*e.g.*, Google) and then mines the returned, top-ranked web pages. These methods depend on the external search engine and requires costly online data extraction.

Later studies, therefore, shift to the *corpus-based* setting, where sets are expanded within a given domain-specific corpus. For example, [6] compute the semantic similarity between two entities based on their local contexts and treat the nearest neighbors around the seed entities as the expanded set. [23] further extend this idea by proposing an iterative similarity aggregation function to calculate entity-entity similarity using query logs and web lists besides free text. More recently, [7, 26] propose to compute semantic similarity using only selected high-quality context features, and [27, 5] develop SetExpander system to leverage multi-context term embedding for entity set expansion. All these attempts, however, assume the input seed entities belong to one unique, clear semantic class, and thus largely suffer from the multi-faceted nature of these seeds – they could represent multiple semantic meanings.

To resolve the ambiguity of seeds, [10] propose to acquire the exact name of the target semantic class and then retrieve its most relevant web tables. Another attempt along this line is EgoSet [11], which utilizes user-created ontology as external knowledge and discovers multiple facets by clustering the expanded entities into different sets. While these semi-structured web tables and ontologies are helpful for disambiguation, they are not always available for domain-specific corpus. More importantly, these two methods are not able to perform quality set expansion when multiple seeds are of different multi-facets (*e.g.* Fig. 1.1). Our proposed FUSE framework only relies on free text and is able to overcome the aforementioned key challenge

of multi-faceted set expansion.

More generally, our work is also related to word sense disambiguation [28, 29, 30, 31, 24]. The major difference is that our work aims to find the coherent semantic facets of all seed words and do entity expansion from the skip-gram clusters.

CHAPTER 6

CONCLUSION

We identify the key challenge of a new problem – *multi-faceted set expansion* and propose a novel approach, FUSE, to address it. By extracting and clustering skip-grams for each seed, identifying coherent semantic facets of all seeds and iteratively expanding entity sets for each semantic facet, FUSE is capable of identifying semantically reasonable facets, generating a quality entity set for each facet, and thus outperforms previous state-of-the-art approaches significantly.

The proposed framework FUSE is general and can incorporate external knowledge for general-domain set expansion. In the future, we plan to explore more on skip-gram representations and quality entity expansion from a set of skip-grams.

REFERENCES

[1] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-aware query suggestion by mining click-through and session data," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[2] P. Velardi, S. Faralli, and R. Navigli, "Ontolearn reloaded: A graph-based algorithm for taxonomy induction," *Computational Linguistics*, vol. 39, no. 3, pp. 665–707, 2013.

[3] J. Zhao, K. Liu, G. Zhou, Z. Qi, Y. Liu, and X. Han, "Knowledge extraction from wikis/bbs/blogs/news web sites," in *Mining User Generated Content*, 2014, pp. 169–206.

[4] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," *Journal of Biomedical Informatics*, vol. 53, pp. 196–207, 2015.

[5] J. Mamou, O. Pereg, M. Wasserblat, A. Eirew, Y. Green, S. Guskin, P. Izsak, and D. Korat, "Term set expansion based NLP architect by intel AI lab," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[6] P. Pantel, E. Crestan, A. Borkovsky, A.-M. Popescu, and V. Vyas, "Web-scale distributional similarity and entity set expansion," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009.

[7] J. Shen, Z. Wu, D. Lei, J. Shang, X. Ren, and J. Han, "Setexpan: Corpus-based set expansion via context feature selection and rank ensemble," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017.

[8] L. Sarmento, V. Jijkoun, M. de Rijke, and E. C. Oliveira, "More like these: Growing entity classes from seeds," in *Proceedings of The Sixteenth ACM Conference on Conference on Information and Knowledge Management*, 2007.

[9] S. Tong and J. Dean, "System and methods for automatically creating lists," 2008, US Patent 7,350,187.

[10] C. Wang, K. Chakrabarti, Y. He, K. Ganjam, Z. Chen, and P. A. Bernstein, "Concept expansion using web tables," in *Proceedings of the 24th International Conference on World Wide Web*, 2015.

[11] X. Rong, Z. Chen, Q. Mei, and E. Adar, "Egoset: Exploiting word ego-networks and user-generated ontology for multifaceted set expansion," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 2016.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013.

[13] M. Steinbach, L. Ertöz, and V. Kumar, "The challenges of clustering high dimensional data," in *New Directions in Statistical Physics*. Springer, 2004, pp. 273–309.

[14] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[15] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[16] K. M. Hermann and P. Blunsom, "Multilingual models for compositional distributed semantics," *arXiv preprint arXiv:1404.4641*, 2014.

[17] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association of Computational Linguistics*, vol. 2, no. 1, pp. 207–218, 2014.

[18] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[19] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[20] M. K. Goldberg, M. Hayvanovych, and M. Magdon-Ismail, "Measuring similarity between sets of overlapping clusters," in *2010 IEEE Second International Conference on Social Computing*, 2010.

[21] N. Chinchor, "Muc-4 evaluation metrics," in *Proceedings of the 4th Conference on Message Understanding*, 1992.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013.

[23] Y. He and D. Xin, "SEISA: Set expansion by iterative similarity aggregation," in *Proceedings of the 20th International Conference on World Wide Web*, 2011.

[24] M. Pelevina, N. Arefyev, C. Biemann, and A. Panchenko, "Making sense of word embeddings," *arXiv preprint arXiv:1708.03390*, 2017.

[25] R. C. Wang and W. W. Cohen, "Language-independent set expansion of named entities using the web," in *Seventh IEEE International Conference on Data Mining*, 2007.

[26] J. Shen, Z. Wu, D. Lei, C. Zhang, X. Ren, M. T. Vanni, B. M. Sadler, and J. Han, "Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

[27] J. Mamou, O. Pereg, M. Wasserblat, I. Dagan, Y. Goldberg, A. Eirew, Y. Green, S. Guskin, P. Izsak, and D. Korat, "Setexpander: End-to-end term set expansion based on multi-context term embeddings," in *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 2018.

[28] K. Taghipour and H. T. Ng, "Semi-supervised word sense disambiguation using word embeddings in general and specific domains," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015.

[29] A. Raganato, C. D. Bovi, and R. Navigli, "Neural sequence learning models for word sense disambiguation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[30] A. Raganato, J. Camacho-Collados, and R. Navigli, "Word sense disambiguation: A unified evaluation framework and empirical comparison," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017.

[31] I. Iacobacci, M. T. Pilehvar, and R. Navigli, "Embeddings for word sense disambiguation: An evaluation study," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.