

Multi-Context Term Embeddings: the Use Case of Corpus-based Term Set Expansion

Jonathan Mamou,¹ Oren Pereg,¹ Moshe Wasserblat,¹ Ido Dagan²

¹Intel AI Lab, Israel

²Department of Computer Science, Bar-Ilan University, Ramat Gan, Israel

¹{jonathan.mamou, oren.pereg, moshe.wasserblat}@intel.com

²dagan@cs.biu.ac.il

Abstract

In this paper, we present a novel algorithm that combines multi-context term embeddings using a neural classifier and we test this approach on the use case of corpus-based term set expansion. In addition, we present a novel and unique dataset for intrinsic evaluation of corpus-based term set expansion algorithms. We show that, over this dataset, our algorithm provides up to 5 mean average precision points over the best baseline.

1 Introduction

Term set expansion is the task of expanding a given seed set of terms into a more complete set of terms that belong to the same semantic class. For example, given a seed of personal assistant application terms like ‘Siri’ and ‘Cortana’, the expanded set is expected to include additional terms such as ‘Amazon Echo’ and ‘Google Now’.

Most prior work on corpus-based term set expansion is based on distributional similarity, where early work is primarily based on using sparse vectors while recent work is based on word embeddings. The prototypical term set expansion methods utilize corpus-based semantic similarity between seed terms and candidate expansion terms. To the best of our knowledge, each of the prior methods used a single context type for embedding generation, and there are no reported comparisons of the effectiveness of embedding different context types. Moreover, the lack of a publicly available dataset hinders the replicability of previous work and method comparison.

In this paper, we investigate the research question of whether embeddings of different context types can complement each other and enhance the performance of computational semantics tasks like term set expansion. To address this question, we propose an approach that combines term embed-

dings over multiple contexts for capturing different aspects of semantic similarity. The algorithm uses 5 different context types, 3 of which were previously proposed for term set expansion and additional two context types that were borrowed from the general distributional similarity literature. We show that combining the different context types yields improved results on term set expansion. In addition to the algorithm, we developed a dataset for intrinsic evaluation of corpus-based set expansion algorithms, which we propose as a basis for future comparisons.

Code, demonstration system, dataset and term embeddings pre-trained models are distributed as part of NLP Architect by Intel AI Lab.¹

2 Related Work

Several works have addressed the term set expansion problem. We focus on corpus-based approaches based on the distributional similarity hypothesis (Harris, 1954). State-of-the-art techniques return the k nearest neighbors around the seed terms as the expanded set, where terms are represented by their co-occurrence or embedding vectors in a training corpus according to different context types, such as linear window context (Pan-tel et al., 2009; Shi et al., 2010; Rong et al., 2016; Zaheer et al., 2017; Gyllensten and Sahlgren, 2018; Zhao et al., 2018), explicit lists (Roark and Charniak, 1998; Sarmiento et al., 2007; He and Xin, 2011), coordinational patterns (Sarmiento et al., 2007) and unary patterns (Rong et al., 2016; Shen et al., 2017). In this work, we generalize coordinational patterns, look at additional context types and combine multiple context-type embeddings.

We did not find any suitable publicly available

¹http://nlp_architect.nervanasys.com/term_set_expansion.html

dataset to train and evaluate our set expansion algorithm. The INEX Entity Ranking track (Demartini et al., 2009) released a dataset for the list completion task. However, it addresses a somewhat different task: in addition to seed terms, an explicit description of the semantic class is supplied as input to the algorithm and is used to define the ground truth expanded set. Some works like (Pantel et al., 2009) provide an evaluation dataset that does not include any training corpus, which is required for comparing corpus-based approaches. Sarmiento et al. (2007) use Wikipedia as training corpus, but exploit meta-information like hyperlinks to identify terms; in our work, we opted for a dataset that matches real-life scenarios where terms have to be automatically identified.

Systems based on our approach are described by (Mamou et al., 2018a,b).

3 Term Representation

Our approach is based on representing any term in a (unlabeled) training corpus by its word embeddings in order to estimate the similarity between seed terms and candidate expansion terms. Different techniques for term extraction are described in detail by Moreno and Redondo (2016). We follow Kageura and Umino (1996) who approximate terms by noun phrases (NPs),² extracting them using an NP chunker. We use *term* to refer to such extracted NP chunk and *unit* to refer to either a term or a word.

As preprocessing, term variations, such as aliases, acronyms and synonyms, which refer to the same entity, are grouped together.³ Next, we use term groups as input elements for embedding training (the remaining corpus words are left intact); this enables obtaining more contextual information compared to using individual terms, thus enhancing embedding model robustness. In the remainder of this paper, by language abuse, *term* will be used instead of term group.

While word2vec originally uses a linear window context around the *focus word*, the literature describes other possible context types. For each *focus unit*, we extract *context units* of different types, as follows (see a typical example for each

²Our algorithm can be used for terms with other part-of-speech or with other term extraction methods.

³For that, we use a heuristic algorithm based on text normalization, abbreviation web resources, edit distance and word2vec similarity. For example, *New York*, *New-York*, *NY*, *NYC* and *New York City* are grouped.

type in Table 1⁴).

3.1 Linear Context (Lin)

This context is defined by neighboring context units within a fixed length window of context units, denoted by *win*, around the focus unit. word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and fastText (Joulin et al., 2016) are state-of-the-art implementations.

3.2 Explicit Lists

Context units consist of terms co-occurring with the focus term in textual lists such as comma separated lists and bullet lists (Roark and Charniak, 1998; Sarmiento et al., 2007).

3.3 Syntactic Dependency Context (Dep)

This context is defined by the syntactic dependency relations in which the focus unit participates (Levy and Goldberg, 2014; MacAvaney and Zeldes, 2018). The context unit is concatenated with the type and the direction of the dependency relation.⁵ This context type has not yet been used for set expansion. However, Levy and Goldberg (2014) showed that it yields more functional similarities of a co-hyponym nature than linear context and thus may be relevant to set expansion.

3.4 Symmetric Patterns (SP)

Context units consist of terms co-occurring with the focus term in symmetric patterns (Schwartz et al., 2015). We follow Davidov and Rappoport (2006) for automatic extraction of SPs from the textual corpus.⁶ For example, the symmetric pattern ‘X rather than Y’ captures certain semantic relatedness between the terms X and Y. This context type generalizes coordinational patterns (‘X and Y’, ‘X or Y’), which have been used for set expansion.

3.5 Unary Patterns (UP)

This context is defined by the unary patterns in which the focus term occurs. Context units con-

⁴We preferred showing in the example the strength of each context type with a good example, rather than providing a common example sentence across all the context types.

⁵Given a focus unit t with modifiers m_i and a head h , the context of t consists of the pairs (m_i/l_i) , where l_i is the type of the dependency relation between the head h and the modifier m_i ; the context stores also (h/l_i^{-1}) where l_i^{-1} marks the inverse-relation between t and h .

⁶SPs are automatically extracted using the dr06 library available at <https://homes.cs.washington.edu/~roysch/software/dr06/dr06.html>.

Cont. type	Example sentence	Context units
Lin $win = 5$	Siri uses voice queries and a natural language user interface.	uses, voice queries, natural language user interface
List	Experience in image processing , signal processing, computer vision.	signal processing, computer vision
Dependency	Turing studied as an undergraduate ... at King's College, Cambridge.	(Turing/nsubj), (undergraduate/prep_as), (King's College/prep_at)
SP	Apple and Orange juice drink ...	Orange
UP	In the U.S. state of Alaska ...	U.S. state of _

Table 1: Examples of extracted context units per context type. Focus units appear in bold.

sist of n -grams of terms and other words, where the focus term occurs; ‘_’ denotes the placeholder of the focus term in Table 1. Following Rong et al. (2016), we extract six n -grams per focus term.⁷

We show in Section 7 that different context types complement each other by capturing different types of semantic relations. As explained in Section 2, to the best of our knowledge, several of these context types have been used for set expansion, except for syntactic dependency context and symmetric patterns. We train a separate term embedding model for each of the 5 context types and thus, for each term, we obtain 5 different vector representations. When training for a certain context type, for each focus unit in the corpus, corresponding $\langle \text{focus unit}, \text{context unit} \rangle$ pairs are extracted from the corpus and are then fed to the `word2vecf` toolkit that can train embeddings on arbitrary contexts, except for linear context for which we use the `word2vec` toolkit. Only terms representations are stored in the embedding models while other word representations are pruned.

4 Multi-Context Seed-Candidate Similarity

For a given context type embedding and a seed term list, we compute two similarity scores between the seed terms and each candidate term, based on cosine similarity.⁸ First, we apply the

⁷Given a sentence fragment $c_{-3} c_{-2} c_{-1} t c_1 c_2 c_3$ where t is the focus term and c_i are the context units, the following n -grams are extracted: $(c_{-3} c_{-2} c_{-1} t c_1)$, $(c_{-2} c_{-1} t c_1 c_2)$, $(c_{-2} c_{-1} t c_1)$, $(c_{-1} t c_1 c_2 c_3)$, $(c_{-1} t c_1 c_2)$, $(c_{-1} t c_1)$.

⁸Sarmento et al. (2007) and Pantel et al. (2009) use first-order semantic similarities for explicit list and coordinational pattern context types, respectively. However, Schwartz et al. (2015) showed that for the symmetric patterns context type, word embeddings similarity (second-order) performs gener-

ally better. We opted for term embeddings similarity (second-order) for all the context types.

⁹Optimal values for k and k' are tuned on the training term list. Other terms are assigned a similarity score of 0 for normalization and combination purpose.

¹⁰For any seed term, cosine similarities are normalized among the candidate terms in order to combine cosine similarity values estimated on different seed terms for the same candidate term, as suggested by Wu et al. (2006).

$centroid$ scoring method ($cent$), commonly used for set expansion (Pantel et al., 2009). The centroid of the seed is represented by the average of the term embedding vectors of the seed terms. Candidate terms become the k terms⁹ that are the most similar, by cosine similarity, to the centroid of the seed. Second, the *CombSUM* scoring method ($csum$) is commonly used in Information Retrieval (Shaw et al., 1994). We first produce a candidate term set for each individual seed term: candidate terms become the k' terms⁹ that are the most similar, according to the term embedding cosine similarity, to the seed term. The CombSUM method scores the similarity of a candidate term to the seed terms by averaging over all the seed terms the normalized pairwise cosine similarities¹⁰ between the candidate term and the seed term.

To combine multi-context embeddings, we follow the general idea of Berant et al. (2012) who train an SVM to combine different similarity score features to learn textual entailment relations. Similarly, we train a Multilayer Perceptron (MLP) binary classifier that predicts whether a candidate term should be part of the expanded set based on 10 similarity scores (considered as input features), using the above 2 different scoring methods for each of the 5 context types. Note that our MLP classifier polynomially combines different semantic similarity estimations and performs better than their linear combination. We also tried to concatenate the multi-context term embeddings in order to obtain a single vector representing all the context

ally better. We opted for term embeddings similarity (second-order) for all the context types.

⁹Optimal values for k and k' are tuned on the training term list. Other terms are assigned a similarity score of 0 for normalization and combination purpose.

¹⁰For any seed term, cosine similarities are normalized among the candidate terms in order to combine cosine similarity values estimated on different seed terms for the same candidate term, as suggested by Wu et al. (2006).

types. We trained an MLP classifier with concatenated vectors of candidate and seed terms as input features, but it performed worst (see Section 7).

5 Dataset

Given the lack of suitable standard dataset for training and testing term set expansion models, we used Wikipedia to develop a standard dataset. Our motivation for using Wikipedia is two-fold. First, Wikipedia contains human-generated lists of terms ('List of' pages) that cover many domains; these lists can be used for supervised training (MLP training in our approach) and for evaluating set expansion algorithms. Second, it contains textual data that can be used for unsupervised training of corpus-based approaches (multi-context term embedding training in our approach). We thus extracted from an English Wikipedia dump a set of term lists and a textual corpus for term embedding training.

5.1 Term Lists

A Wikipedia 'List of' page contains terms belonging to a specific class, where a term is defined to be the title of a Wikipedia article. We selected term lists among 'List of' pages containing between fifty and eight hundred terms in order to cover both specific and more common classes (e.g., list of *chemical elements* vs. list of *countries*). Moreover, we selected term lists that define purely a semantic class, with no additional constraints (e.g., skipping list of *biblical names starting with N*). Since there can be some problems with some Wikipedia 'List of' pages, 28 term lists have been validated manually and are used as ground truth in the evaluation. Here are some few examples of term lists: Australian cities, chemical elements, countries, diplomatic missions of the United Kingdom, English cities in the United Kingdom, English-language poets, Formula One drivers, French artists, Greek mythological figures, islands of Greece, male tennis players, Mexican singers, oil exploration and production companies.

Terms having a frequency lower than 10 in the training corpus are pruned from the lists since their embeddings cannot be learned properly; note that these terms are generally less interesting in most of real case applications. Term variations are grouped according to Wikipedia redirect information.

On average, a term list contains 328 terms, of which 3% are not recognized by the noun phrase chunker; the average frequency of the terms in the corpus is 2475.

The set of term lists is split into train, development (dev) and test sets with respectively 5, 5 and 18 lists for MLP training, hyperparameters tuning and evaluation. Each term list is randomly split into *seed* and *expanded* term sets, where we are interested in getting enough samples of seed and expanded term sets. Thus, given a term list, we randomly generate 15 seed sets (5 seed sets for each seed size of 2, 5 and 10 terms) where seed terms are sampled among the top 30 most frequent terms within the list. For the train set, the non-seed terms (expanded term set) provide the positive samples; we randomly select candidate terms that occur in the corpus but not in the list as negative samples; positive and negative classes are balanced.

5.2 Textual Corpus

The corpus contains all the textual parts of Wikipedia articles except 'List of' pages.¹¹ It is used for training the multi-context embedding models. 3% of the terms appearing in the term lists are not recognized by our NP chunker in the corpus. It contains 2.2 billion words, and 12 million unique terms are automatically extracted.

5.3 Public Release

We use `enwiki-20171201` English Wikipedia dump¹² to develop the dataset. Full dataset will be released upon publication and it will include train, dev and test sets including the split into seed and expanded terms, and negative samples for the train set; the textual corpus along with NP chunks and grouped term variations; term embedding model for each context type.

6 Implementation Details

Code is distributed under the Apache license as part of NLP Architect by Intel AI Lab¹³, an open-source Python library for exploring state-of-the-art deep learning topologies and techniques for natural language processing and natural language understanding.

¹¹Note that the corpus does not contain any Wikipedia meta information.

¹²[enwiki-20171201-pages-articles-multistream.xml.bz2](https://s3.amazonaws.com/wiki-dumps/enwiki-20171201-pages-articles-multistream.xml.bz2)

¹³http://nlp_architect.nervanasys.com

We used the following tools for the implementation and for the development of the dataset: `spaCy`¹⁴ for tokenization, noun phrase chunking and dependency parsing; `textacy`¹⁵ for text normalization; `word2vec`¹⁶ and `fastText`¹⁷ to model term embeddings of linear context type; `word2vecf`¹⁸ to model term embeddings of other context types; `WikiExtractor`¹⁹ to extract textual part of Wikipedia dump; `Keras`²⁰ to implement the MLP classifier.

Similarity scores are softmax-normalized over all the candidate terms per context type and per scoring method, in order to combine them with the MLP classifier. Our MLP network consists of one hidden layer. The input and hidden layers have respectively ten and four neurons.

7 Experiments

Following previous work (Sarmiento et al., 2007), we report the Mean Average Precision at several top n values (MAP@ n) to evaluate ranked candidate lists returned by the algorithm. When computing MAP, a candidate term is considered as matching a gold term if they both appear in the same term variations group. We first compare the different context types; then, we report results on their combination.

7.1 Context Type Analysis

We provide a comparison of the different context types in Table 2. These context types are baselines and we compare them to the linear context that is more standard. Note that the dependency context type is affected by the performance of the dependency parser.²¹ Linear context with centroid scoring yields consistently best performance of at least 19 MAP@10 points and is consistently more stable looking at standard deviation. However, other context types achieve better performance than linear context type for 55% of the term lists, suggesting that the different context types complement each other by capturing better different types of

¹⁴<https://spacy.io>

¹⁵<https://github.com/chartbeat-labs/textacy>

¹⁶<https://code.google.com/archive/p/word2vec>

¹⁷<https://github.com/facebookresearch/fastText>

¹⁸<https://bitbucket.org/yoavgo/word2vecf>

¹⁹<https://github.com/attardi/wikiextractor>

²⁰<https://github.com/keras-team/keras>

²¹We used `spaCy` for dependency parsing; it achieves 92.6% accuracy on the OntoNotes 5 corpus (Choi et al., 2015).

Context	Scor.	MAP@10	stdev	best %
Lin	cent	.78	.22	
List	csum	.59	.30	20
Dep	cent	.53	.31	15
SP	csum	.48	.32	10
UP	csum	.47	.36	10

Table 2: Comparison of the different context types. For each context type, we report the scoring method with higher MAP@10 on dev set, MAP@10 with 5 seed terms, its standard deviation among the different test term lists, the percentage of the test term lists where the context type achieves best performance.

Method	MAP@10	MAP@20	MAP@50
Linear	.78	.71	.59
Concat.	.68	.65	.56
MLP	.83	.74	.63
Oracle	.89	.82	.73

Table 3: MAP@ n performance evaluation of the linear context, concatenation, MLP binary classification and oracle, with 5 seed terms.

semantic relations and that their combination may improve the quality of the expanded set.

In addition, performance consistently increases with the number of seed terms e.g., MAP@10, MAP@20 and MAP@50 of the linear context are respectively .66, .58 and .51 with 2 seed terms.

7.2 Context Combination

We provide in Table 3 MAP@ n for the centroid scoring of the linear context and for the MLP classification with 5 seed terms. For comparison, we report in ‘Concat.’ row the performance for the MLP binary classification on the concatenation of the multi-context term embeddings. In addition, we report *oracle* performance assuming we have an oracle that chooses, for each term list, the best context type with the best scoring method. Oracle performance shows that the context types are indeed complementary. The MLP classifier which combines all the context types, yields additional improvement in the MAP@ n compared to the baseline linear context. Moreover, we observed that the improvement of the MLP combination over the linear context is preserved with 2 and 10 seed terms. Yet, looking at the oracle, the MLP combination still does not optimally integrate all the information captured by the term embeddings.

8 Conclusion

We proposed a novel approach to combine different context embedding types and we showed that it achieved improved results for the corpus-based term set expansion use case. In addition, we publish a dataset and a companion corpus that enable comparability and replicability of work in this field.

For future work, we plan to run similar experiments using recently introduced contextual embeddings, (e.g., ELMo (Peters et al., 2018), BERT (Devlin et al., 2018), OpenAI GPT-2 (Radford et al., 2019)), which are expected to implicitly capture more syntax than context-free embeddings used in the current paper. We plan also to investigate the contribution of multi-context term embeddings to other tasks in computational semantics.

References

- J. Berant, I. Dagan, and J. Goldberger. 2012. Learning entailment relations by global graph structure optimization. *Computational Linguistics*, 38:73–111.
- Jinho D Choi, Joel Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 387–396.
- Dmitry Davidov and Ari Rappoport. 2006. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 297–304. Association for Computational Linguistics.
- Gianluca Demartini, Tereza Iofciu, and Arjen P De Vries. 2009. Overview of the inex 2009 entity ranking track. In *International Workshop of the Initiative for the Evaluation of XML Retrieval*, pages 254–264. Springer.
- 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*.
- Amaru Cuba Gyllensten and Magnus Sahlgren. 2018. Distributional term set expansion. *arXiv preprint arXiv:1802.05014*.
- Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.
- Yeye He and Dong Xin. 2011. Seisa: set expansion by iterative similarity aggregation. In *Proceedings of the 20th international conference on World wide web*, pages 427–436. ACM.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, H erve J egou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Kyo Kageura and Bin Umno. 1996. Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2):259–289.
- Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 302–308.
- Sean MacAvaney and Amir Zeldes. 2018. A deeper look into dependency-based word embeddings. *arXiv preprint arXiv:1804.05972*.
- Jonathan Mamou, Oren Pereg, Moshe Wasserblat, Ido Dagan, Yoav Goldberg, Alon Eirew, Yael Green, Shira Guskin, Peter Izsak, and Daniel Korat. 2018a. Term Set Expansion based on Multi-Context Term Embeddings: an End-to-end Workflow. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*.
- Jonathan Mamou, Oren Pereg, Moshe Wasserblat, Alon Eirew, Yael Green, Shira Guskin, Peter Izsak, and Daniel Korat. 2018b. Term Set Expansion based NLP Architect by Intel AI Lab. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Antonio Moreno and Te ofilo Redondo. 2016. Text analytics: the convergence of big data and artificial intelligence. *IJIMAI*, 3(6):57–64.
- Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu, and Vishnu Vyas. 2009. Web-scale distributional similarity and entity set expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 938–947. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Brian Roark and Eugene Charniak. 1998. Noun-phrase co-occurrence statistics for semiautomatic semantic lexicon construction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 1110–1116. Association for Computational Linguistics.
- Xin Rong, Zhe Chen, Qiaozhu Mei, and Eytan Adar. 2016. 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*, pages 645–654. ACM.
- Luis Sarmiento, Valentin Jijkuon, Maarten de Rijke, and Eugenio Oliveira. 2007. More like these: growing entity classes from seeds. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 959–962. ACM.
- Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 258–267.
- Joseph A. Shaw, Edward A. Fox, Joseph A. Shaw, and Edward A. Fox. 1994. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–252.
- 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 *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 288–304. Springer.
- Shuming Shi, Huibin Zhang, Xiaojie Yuan, and Ji-Rong Wen. 2010. Corpus-based semantic class mining: distributional vs. pattern-based approaches. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 993–1001. Association for Computational Linguistics.
- Shengli Wu, Fabio Crestani, and Yaxin Bi. 2006. Evaluating score normalization methods in data fusion. In *Asia Information Retrieval Symposium*, pages 642–648. Springer.
- Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. 2017. Deep sets. In *Advances in Neural Information Processing Systems*, pages 3394–3404.
- He Zhao, Chong Feng, Zhunchen Luo, and Chang-hai Tian. 2018. Entity set expansion from twitter. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR '18*, pages 155–162, New York, NY, USA. ACM.