Memory, Show the Way: Memory Based Few Shot Word Representation Learning

Jingyuan Sun[1,2], **Shaonan Wang**[1,2], **Chengqing Zong**[1,2,3]
[1] National Laboratory of Pattern Recognition, CASIA, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{jingyuan.sun, shaonan.wang,cqzong}@nlpr.ia.ac.cn

Abstract

Distributional semantic models (DSMs) generally require sufficient examples for a word to learn a high quality representation. This is in stark contrast with human who can guess the meaning of a word from one or a few referents only. In this paper, we propose Mem2Vec, a memory based embedding learning method capable of acquiring high quality word representations from fairly limited context. Our method directly adapts the representations produced by a DSM with a longterm memory to guide its guess of a novel word. Based on a pre-trained embedding space, the proposed method delivers impressive performance on two challenging few-shot word similarity tasks. Embeddings learned with our method also lead to considerable improvements over strong baselines on NER and sentiment classification.

1 Introduction

Humans can learn a new word quickly from minimal exposure to its context, as in the following example:

> The *Labrador* runs happily towards me, barking and wagging its tail.

Even this is the first time one hears about *Labrador*, we can guess it should be an animal or even further a dog easily, since it runs, barks and has a tail. Such ability to efficiently acquire representation from small data, namely fast mapping, is thought to be the hallmark of human intelligence that a cognitive plausible agent should strive to reach (Xu and Tenenbaum, 2007; Lake et al., 2015).

However, as the mainstream of text representation learning in NLP, most distributed semantic models (DSMs) don't fare well in tiny data (Lazaridou et al., 2017; Herbelot and Baroni, 2017; Wang et al., 2016). Even if they have learned a lot of words, they still need sufficient examples to acquire a high-quality representation for a novel word. This not only constitutes a blow to DSM's cognitive plausibility but also limits its practical usage in NLP. Because plentiful enough data is not always available, especially in domain specific tasks. Even if a large corpora is at hand, low-frequency words in it are still more than highly frequent ones, according to the Zipfian distribution of natural language.

Given the above reasons, it's desirable to build a word embedding method capable of acquiring high quality representations with limited contexts, i.e., few shot word representation learning. We take lessons from hypothesis constraint (HC) theory to achieve this goal. HC is an influential proposal for human's fast mapping (Xu and Tenenbaum, 2007). It indicates that people learn a new word by eliminating incorrect hypotheses about the word meaning, based on usage of the target word and prior knowledge of context words. This is instructive to us since embedding a word in the high-dimensional vector space also faces nearly unlimited candidate hypotheses (Wang et al., 2018) . General DSMs can't efficiently handle these candidates, so they fall back on multiple context examples to find the path, while we propose to let a memory show the way. We augment DSM with an longterm memory to transfer knowledge from a large general domain corpora to adapt the representation learning on the small text. In context of the HC theory, we directly constrain the hypothesis a DSM makes about the target word by its current usage and prior knowledge acquired from a large corpora. Experiments show our method makes educated guess of a novel word efficiently with fairly limited examples, just as humans do in the fast mapping.

It's worth noting that us attaching importance

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1435–1444
Brussels, Belgium, October 31 - November 4, 2018.

to few-shot word representation learning doesn't mean we need to learn words, no matter frequent or rare, all in the few-shot way. DSMs have done pretty well in frequent words learning with large corpus. We augment a DSM with an external memory for few-shot representation learning, under the assumption that gradual acquisition of frequent words plus fast learning of rare words make an integrated word representation learning scheme. Our ultimate goal is certainly an all-round architecture that learns text representations from any amount of data. We believe Mem2Vec, which bridges the word representation learning from big data to small text, will be a building block of that ideal architecture.

The primary contribution of this work is a memory augmented word embedding model with a fast adaptation mechanism, capable of learning representations efficiently from tiny data. Experimental results show that the proposed Mem2Vec learns high quality target word representation with both single informative sentence and a few casual sentences as contexts. To show its performance in downstream applications, Mem2Vec is used to pre-train embeddings for three NER tasks and also surpasses strong baselines. Since our model transfers from general domain corpus to a target small text, it is highly possible to face the problem of domain shift. Mem2vec is impressively competent in tackling domain shift, as demonstrated in a series of cross-domain sentiment classification tasks.

2 Related Work

Rare Word Embedding Acquiring representations for rare words has long been a well-known challenge of natural language understanding (Herbelot and Baroni, 2017; Wang and Zong, 2017). Khodak et al. (2018) learn a linear transformation with pretrained word vectors and linear regression, which can be efficient adapted for novel words. Lazaridou et al. (2017) directly sum the context embedding of a novel word as its representation, based on a pre-trained embedding space. Though not explicitly stated, their idea actually matches the HC theory. They constrain the hypothesis solely within the current context of the target word which we think is not enough. We constrain the hypothesis with memory and the context. Another strand of solutions rely on auxiliary information, such as morphological structure (Luong et al., 2013; Kisselew et al., 2015) and external knowledge (Long et al., 2017). Lazaridou et al. (2013) derive morphologically complex words from sub-word parts with phrase composition methods. Ling et al. (2015) read characters of the rare word with a bidirectional LSTM to deal with open vocabulary problem in language modeling and NER. Hill et al. (2016) learn an embedding of a dictionary definition to match the pre-trained headword vector, while Weissenborn (2017) refines the word embeddings with explicit background knowledge from a commonsense knowledge base. Different from this strand of work, our method doesn't fall back on auxiliary information. We acquire knowledge from a large unlabeled general domain corpora which is widely available.

Cross Domain Word Embedding The knowledge accumulation phase of our model aims to learn an embedding space from a large general domain corpora. This is partially in line with cross domain word embedding work. Among these work, a strand of approach hypothesizes that a word frequent in multiple domains should mean nearly across these domains. Bollegala et al. (2015) call such word pivot, share its embeddings across domains and use them to predict the surrounding non-pivots. Yang et al. (2017) selectively incorporate source domain information to target domain word embeddings with a word-frequency-based regularization. These pivot-based methods have delivered improvements on sentiment analysis and NER. However, they have a defect that only limited target domain words benefit from the knowledge transfer.

Memory based Meta Learning Memory augmented neural networks (MANN) are widely used in different tasks for efficient recall of experience and fast adaptation to new knowledge (Bahdanau et al., 2014; Merity et al., 2017; Miller et al., 2016; Grave et al., 2017; Sprechmann et al., 2018; Wang et al., 2017). Intuitively, Meta-learning, which aims to train a model that quickly adapts to a new task, should benefit from memory architecture, and empirically it does do (Santoro et al., 2016; Duan et al., 2016; Wang et al., 2016; Munkhdalai and Yu, 2017; Kaiser et al., 2017). The memory we use is closely related to (Kaiser et al., 2017), but still get three major differences. First,they only retrieve the single nearest neighbor from the memory while we retrieve an average of the K nearest neighbors weighted by how they match the current context. Second, they focus on supervised learn-

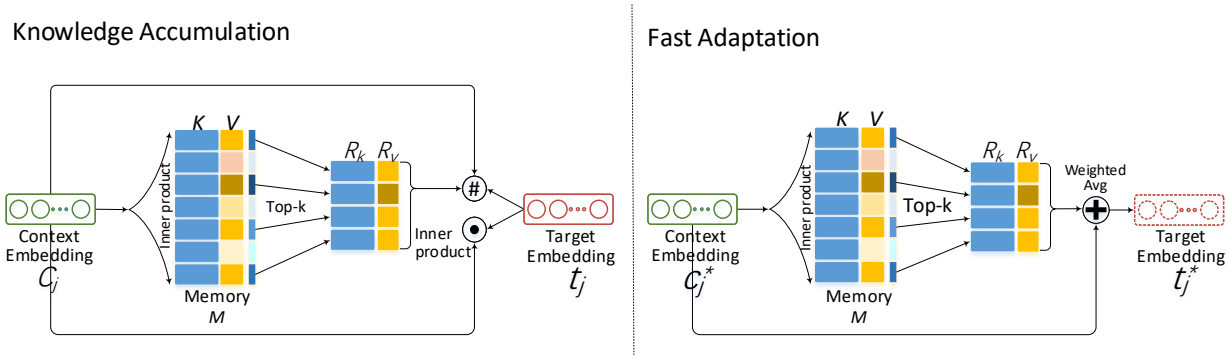

Figure 1: The proposed model architecture. *K* and *V* respectively denotes the key and value vector of the memory. # refers to equation(3). Left: Knowledge accumulation phase. The model learns word embeddings and store prototypes in memory. Right: Fast Adaptation phase. Prototypes retrieved from memory with the given context are combined with the context embedding to form the target word representation.

ing and don't have a fast adaptation mechanism for acquiring representation. Third, they update the memory according to whether the returned value is strictly the same as the target. However, synonyms are common in natural language text. We thus take a softer criterion and update the memory according to vector similarity between the addressed value and the target word embedding.

3 Methods

Our model in brief is a neural network based DSM augmented by a longterm memory. As showed in Fig.1, it operates in two consecutive phases, first accumulating knowledge and then doing fast adaptation on new words, just as the human learning process goes. In the knowledge acquisition phase, we train the memory augmented DSM to learn a semantic space. We also accumulate similar contexts of target words in the memory and gradually form "prototype" representations. The pre-trained embedding space and the saved prototypes are just the knowledge acquired. The fast adaption phase occurs when we need to learn a new word from minimal context. In this phase, we directly combine the context embedding and retrieve content from the memory to form the target word representation.

In the following sections, we will first introduce the memory architecture and the content based addressing. We then detail how exactly our model runs respectively in the knowledge accumulation and fast adaption phase.

3.1 Memory Addressing

M is a non-parametric key-value memory which stores a key-value pair (k_i, v_i) in each memory slot i. Inspired by (Kaiser et al., 2017), we keep an additional vector A tracking the age of slots. The initial age of all is zero. So the whole memory M looks like $(K_{m\times ks}, V_m, A_m)$ where m denotes memory size and ks denotes key vector size. Given a normalized query q, its nearest neighbor in M is defined as any of the keys that maximize the cosine similarity with q:

$$\mathrm{NN}(q, M) = \arg\max_i \ q \cdot K_i. \tag{1}$$

During training, a query to the memory M searches k nearest neighbors which is a natural extension to (1):

$$(n_1, ..., n_k) = \mathrm{NN_k}(q, M). \tag{2}$$

We take an average weighted by how the addressed memory slots match the query:

$$R_K, R_V = \sum_{k=i}^{K} \mathrm{softmax}(\frac{q \cdot M[n_k]}{\sqrt{d_k}}) \cdot M[n_k]. \tag{3}$$

This is actually a dot-product attention on the k nearest neighbors. R_K, R_V are the final output of the memory. Note that here we use softmax with temperature T:

$$softmax(a) = \frac{\frac{e^a}{T}}{\Sigma_{i=1}^{n} \frac{e^i}{T}}. \tag{4}$$

T is normally set to 1. Using a higher value for T produces a softer probability distribution. We

set different temperatures in the knowledge accumulation and fast adaptation phase, which will be detailed in the following subsections.

3.2 Knowledge Accumulation

Given a target word embedding t_j with its context embedding c_j as input, we query the memory with c_j as (1)-(3) and retrieve R_k, R_V. The semantic relation between the current example and the retrieved content from memory is hoped to be consistent from context to target words, so we derive the following loss:

$$\mathcal{L}_m = \sum_{t_j, c_j \in D} log\sigma\left((R_K - R_V) \cdot (c_j - t_j)\right) \quad (5)$$

We also hope the target word representation fully incorporates context information and stay far from negative examples, so we also inherit the loss from (Mikolov et al., 2013):

$$\begin{aligned}\mathcal{L}_s = \sum_{t_j, c_j \in D} \#(t_j, c_j)\Big(log\sigma(t_j \cdot c_j) \\ + \sum_{i=1}^{k} \mathbb{E}_{p_i \sim P(t_j)}[log\sigma\ (t_j \cdot -p_i)]\Big),\end{aligned} \quad (6)$$

where D denotes the whole corpora, and $\#(t_j, c_j)$ means times the target and the context word co-occur. The word p_i is a negative sample sampled from the distribution $P(t_j)$, as Mikolov et al. (2013) do. We minimize the sum of $\mathcal{L}_m$ and $\mathcal{L}_s$:

$$\mathcal{L} = \mathcal{L}_m\ + \mathcal{L}_s \quad (7)$$

Memory Update. After each query, we update a memory slot according to how frequently the key is addressed and how useful the addressed value is. The update is done piecewise according to similarity between the addressed values $(V_{n_1}, ..., V_{n_k})$ and the target word. For all the addressed values V_{n_i} whose similarity to the target word is higher than the threshold β, we only update its corresponding key by taking a weighted average of the current key and the query:

$$K[n_i] \leftarrow \frac{q + K[n_i]}{||q + K[n_i]||}. \quad (8)$$

Otherwise, it means no addressed value correlates enough with the target, we then choose memory slots n' with maximum age and rewrite the stored items in it:

$$K[n'] \leftarrow q, V[n'] \leftarrow t_j. \quad (9)$$

The age of each updated slot will be reset to zero while all other non-updated slots get incremented by 1 in age. Memory updated in this way gradually accumulates similar contexts of a word into the same slot, which in another word, forms the prototype representation of a word.

3.3 Fast Adaptation

Now we show how to poll the memory to efficiently learn a new word representation from limited context. This is where the hypothesis constraint takes place. To be specific, given a new word embedding t_j^* to be learned and its context embedding c_j^*, we retrieve memory relevant to c_j^* as (2)-(3) and get R_K^*. Then we adapt context embedding with the retrieved memory to form the target word representation:

$$t_j^* = \alpha R_K^* + (1 - \alpha) c_j^*, \quad (10)$$

where α can be tuned a hyper-parameter or learned with regression models. Actually we also try to incorporate R_V^*, but the aggregated prototype R_K^* seems to continuously perform better.

We here pay additional attention to the softmax temperature T. T is emphasized since it conditions how the model "treats" the retrieved memory. Contexts are fairly limited in the few-shot case, so how the retrieved memory is treated crucially affects the quality of the learned representation. A higher temperature leads to a softer attention distribution , which means the model will be more likely to sample from all retrieved contents. A lower temperature makes the model focus more on the memory with highest similarity to the query. We predict a slightly higher temperature will generally be better in the fast adaptation phase. Since the HC theory points out that hypotheses are not in strict mutual-exclusions, they overlap with each other which corresponds to the higher-temperature condition. We will test this in the experiments.

4 Few-shot Word Similarity Tasks

We test the proposed method on two few-shot word similarity tasks. Fig.2 gives examples of the two tasks. In the following subsections we will introduce these datasets in detail and show the performance of tested methods on these tasks.

4.1 Tasks and Datasets

Definitional Nonce Task We evaluate on the Definitional Nonce dataset (Herbelot and Baroni,

Nonce Definition
Provided Context :
___ international inc is an american multinational conglomerate company that produces a variety of commercial and consumer products engineering services and aerospace systems for a wide variety of customers from private consumers to major corporations and governments
Ground Truth Word: Honeywell

Chimera-l2
Provided Context :
1. Canned sardines and ____ between two slices of whole meal bread and thinly spread Flora Original.
2. Erm, ____, low fat dairy products, incidents of heart disease for those who have an olive oil rich diet.

Probe Words: rhubarb, onion, pear, strawberry, limousine, cushion
Human Response: 2.57, 4.43 , 3.86, 3.71, 1.43, 2.14

Figure 2: Examples of the Nonce Definition and Chimera Task

2017) to simulate the process where a competent speaker learns a novel word from one informative sentence. 1000 words are included in the dataset as targets, with 700 for training and 300 for testing. Each target word corresponds to only one sentence extracted from its Wikipedia definition as context. All context sentences have been manually checked to be definitional enough to describe the corresponding target words. After tuning parameters on training data, the model is required to learn the target word representation with the provided context in test set. Learned representations are assessed by similarity to ground truth vectors produced in exposure to the whole corpora. We use the Reciprocal Rank (RR) of the ground vector in all nearest neighbors to the learnt representation for fair comparison of different methods, following Herbelot and Baroni (2017)'s settings. The mean value of RR over all test instances in the dataset is calculated as the final score.

Chimera Task Our second evaluation on the Chimera dataset (Lazaridou et al., 2017) means to simulate the case where a speaker learns the new word in a more casual multi-sentence context, not as highly informative as definitions in the Nonce dataset. There are 3 sub-tasks in Chimera:L2, L4 and L6, respectively providing 2, 4, 6 sentences as context to for each of the 330 instances in the dataset. The tested model needs to learn target word representation from the provided contexts. The similarity between learned embeddings and each of the probe words is measured and compared to human judgments by Spearman correlation. The final score is the average Spearman across all test pairs.

4.2 Baselines

Our model is compared to several baselines, including Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), SUM (Lazaridou et al., 2017) and N2V (Herbelot and Baroni, 2017). Glove and Word2Vec are representatives of traditional DSMs. With them we want to test how exactly traditional DSMs perform in the few shot representation learning without any additional mechanism for small data. SUM and N2V are proposed especially for learning on small corpus. They adapt Word2Vec's skip-gram structure for incremental learning and show improvements on the Chimera dataset. They partially match the HC theory which Mem2Vec is based on. Note that several rare word learning methods (Long et al., 2017; Xu et al., 2014; Lazaridou et al., 2013) that rely on auxiliary information don't apply with most of our task settings. In the Nonce and Chimera task, context for target word learning is strictly limited for fair comparison, so external knowledge is banned. And the target word, as showed in Fig.2, is just a slot which doesn't provide any morphological hints, so sub-word methods are also excluded.

Both the above baselines and the proposed Mem2Vec use a dump of Wikipedia to learn a fundamental semantic space. To be specific, N2V and SUM use embeddings pre-trained by Word2Vec, while Mem2Vec acquires prior-knowledge, all from that Wiki corpora. We calculate correlation with the similarity ratings in the MEN and SIMLEX dataset to verify if the pre-trained semantic space is ready for use.

Model \ Task	Nonce		Chimera		
	MRR	Med. Rank	L2 ρ	L4 ρ	L6 ρ
Word2Vec	0.00007	111012	0.1459	0.2457	0.2498
GloVe	0.00008	108002	0.1402	0.2397	0.2533
SUM	0.03686	861	0.3176	0.3534	0.3880
N2V	0.04907	623	0.3120	0.3628	0.3790
Mem2Vec	**0.05416**	**518**	**0.3301**	**0.3717**	**0.3897**

Table 1: Results of Nonce Definition and Chimera task. MRR and Med Rank respectively denotes mean reciprocal rank and median rank of the ground truth word. ρ denotes precision.

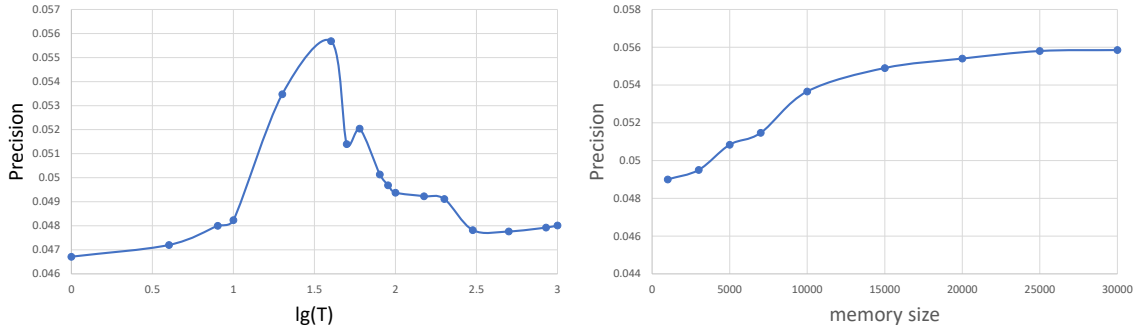

Figure 3: Performance of Nonce Definitional Task under different softmax temperature (left) and memory size (right). The left figure is in semilog coordinate.

4.3 Results

Nonce Definition Task We show the results of Nonce Definition Task in Table 1. Before analyzing the results we need to explain that MRR achieved by the tested models seems pretty low. This is no odd since matching the ground truth word in a vocabulary of 210,512 sets a potentially very large denominator in the reciprocal rank calculation. Our model achieves an MRR of 0.05416, which means the median rank of the true vector is 518 in the challenging two hundred thousand neighbors, surpassing all the baselines. N2V and SUM also deliver satisfactory performance with N2V working better. We are sorry to find that the naive Word2Vec and GloVe totally fail in the Nonce task, supporting the importance of adapting traditional DSMs for few-shot word representation learning.

Chimera Task The results on 3 chimera tasks are shown in Table 1, too. Mem2vec out-performs baselines in all the 3 context length settings. SUM performs steadily well from Nonce to the Chimera task, suggesting the effectiveness of constraining hypothesis space with contexts. But the continuous improvement of Mem2Vec over SUM confirms the advantage of our model, which incorporates "global" semantic information from the memory with the local contexts. N2V also works here but not as well as in the Nonce task, probably because the contexts in chimera are not as informative. Such performance drop may indicate N2V's limited scalability to downstream NLP tasks since not all real world texts are as informative as in Nonce Definition Task. We will test this speculation with NER tasks in section 5.

4.4 Memory Parameter Analysis

We are interested to know how the two key parameters of memory, the softmax temperature T, and the memory size influence the quality of learned representations. We use the Nonce Definition task as the testbed. While studying the influence of one parameter, the other parameters are fixed. We run the model for 3 times with each candidate parameter and calculate the average precision as the final score.

Softmax Temperature Fig.3 (left) shows task performance under different softmax temperatures in semilog coordinate. We are a little bit surprised to find that it roughly fits a normal distribution and a mid-high temperature leads to best performance. A mid-high T means the model doesn't give too large or too small weights to certain retrieved memory. This meets the HC theory about

how humans weight the constrained hypothesis. We don't just trust a single hypothesis, nor do we treat all the hypotheses equally. The experiments shows similar principle also applies to our model.

Memory Size Fig.3 (right) shows task performance under different memory size. We find that increasing the memory size does lead to improved performance . But the improvements tend to be minor after the memory size is larger than 20,000. We owe it to the fact that we does not save specific examples, we accumulate similar contexts together in the memory to form prototypes. While we retrieve the memory, prototypes can be combined in different ways to represent multiple examples, thus a smaller memory can also work as well as the bigger one.

5 Extrinsic Tasks

We hope that the learned representations not only perform well on word similarity tasks but also apply to downstream NLP tasks. NER on domain specific datasets is an ideal benchmark. Named entities in these datasets are relatively low in frequency and not well covered by general domain corpus, tough for a traditional DSM to learn. Besides, while transferring from general domain corpus to a target small text, domain shift is a highly possible issue. We test if Mem2Vec could tackle the domain shift with a series of cross-domain sentiment analysis tasks.

5.1 Tasks and datasets

Domain Specific NER We use BioNLP11-species (Kim et al., 2011), AnatEMs (Pyysalo and Ananiadou, 2013) and NCBI-disease (Doğan et al., 2014) dataset, respectively from taxonomy, anatomy and pathology literatures. We train embeddings with tested methods to initialize the recognizer, whose performance then demonstrates whether the tested models learn representations well for rare words.

Cross Domain Sentiment Classification cross domain sentiment classification on Amazon Review dataset (Blitzer et al., 2007) is chosen as a benchmark. This dataset includes reviews from 4 product categories: books, DVDs, kitchens and electronics, suitable for the cross-domain setting. Using one as source domain and one as the target, we get 16 pairs for experiments. We train the classifier with source domain data and directly test it on the target domain, using the pre-trained embeddings as input feature. Note that through this task we also test how Mem2Vec performs when transferring from a small text, since in all the above experiments we learn prior knowledge from a large corpora.

5.2 Baselines

Except for the four baselines considered in word similarity tasks, we also compare with DAREP (Bollegala et al., 2015) and CRE (Yang et al., 2017) in the NER and sentiment classification tasks. They are both pivot-based methods for cross domain embedding learning which fare well in some downstream tasks. Besides we introduce SCL(Blitzer et al., 2006), a well-cited cross-domain sentiment analyser, as a baseline only for the sentiment classification task.

For NER, we use pre-trained embeddings by the tested methods as only input features for a LSTM-CRF recognition model (Lample et al., 2016). We simply mix the Wikipedia corpora with a dump of PubMed as our source corpora. Note that N2V and SUM can't be directly used to pre-train embeddings for downstream tasks since they focus on novel word learning. We thus explicitly divide words which occur less than 5 times as rare words while others as frequent words. N2V and SUM learn the frequent words with Word2Vec and learn the rare words in their own way. This setting also applies to the sentiment classification task.

For sentiment classification, we use a multi-layer perceptron (MLP) as the classifier, with one hidden layer of 400 nodes, ReLu activation and softmax output function.

5.3 Results

Named Entity Recognition Table 2 shows the results of domain specific named entity recognition. Used for pre-training embeddings, Mem2Vec achieves higher F1-score than all the baselines. It first surpasses CRE and DAREP that only bring slight improvements over Word2Vec. CRE and DAREP are both methods which relies on words with cooccurance patterns in source and target domain as the pivots for cross-domain transfer. This indicates the advantage of Mem2Vec over the traditional word frequency based methods in fast mapping cases where word cooccurrence pattern is not clear.

Our improvements over the N2V and SUM are more obvious than in the two word similarity

Task / Model	AnatEM			BioNLP			NCBI		
	P	R	F1	P	R	F1	P	R	F1
Word2Vec	76.12	69.80	72.82	73.13	54.79	62.64	75.22	75.37	74.39
GloVe	75.83	67.04	71.14	72.58	53.35	61.50	75.76	72.33	74.01
N2V	76.81	66.8	71.46	73.91	54.21	62.54	72.45	74.37	73.30
SUM	77.06	69.01	72.81	74.36	58.58	62.25	74.89	74.02	74.45
DAREP	79.03	**67.95**	73.07	**77.18**	54.19	63.67	78.76	75.60	77.15
CRE	80.04	67.90	73.47	76.74	56.98	65.40	78.98	**76.63**	77.79
Mem2Vec	**81.23**	67.90	**73.96**	76.70	**57.81**	**65.92**	**79.56**	**76.63**	**78.06**

Table 2: Results of domain specific Named Entity Recognition. P, R, F1 respectively denotes precision, recall and F1 score

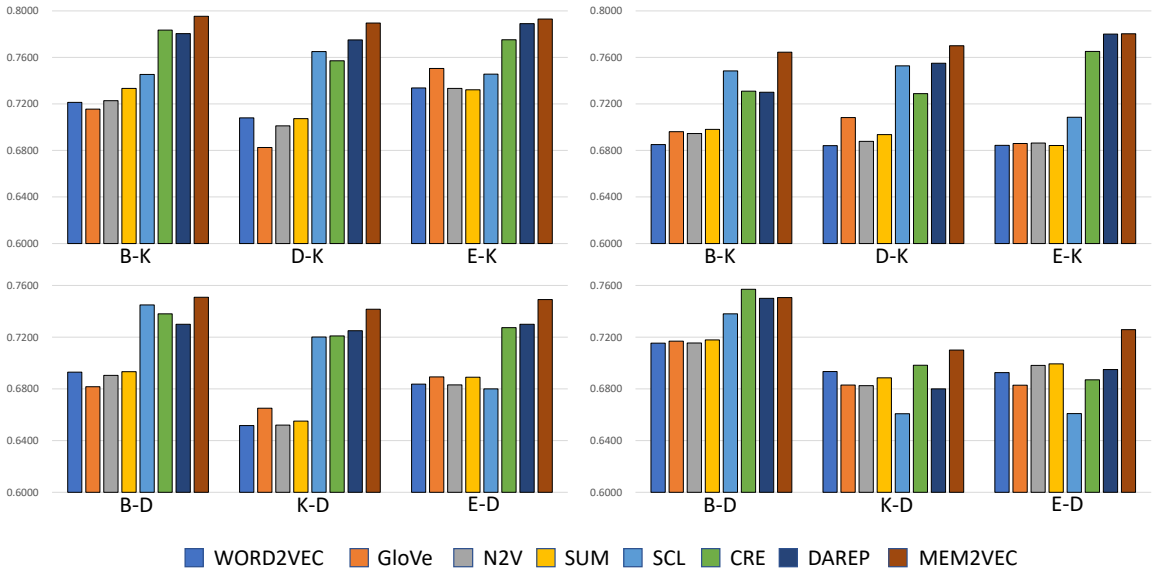

Figure 4: Results of cross domain sentiment classification on Amazon Review dataset. B denotes books, D for DVD, E for electronics, K for kitchen. B-K means B is the source domain and K is the target domain.

tasks. This again affirms our speculation that constraining hypothesis solely with the context is not enough. In the setting of NER, the context of one named entity is likely to be filled with other named entities which are also low in frequency. Directly summing the context as SUM does or taking risk to enlarge the window size as N2V may lead to over-fitting. While every training step of our method incorporates relative information from all the experienced examples stored in the memory, alleviating the danger of learn representations that over fits the local contexts.

In addition, it's worth noting that parameter tuning for N2V is no picnic. In our experiments, the original settings: high learning rate, large window size and short iteration span don't lead to satisfactory performance. More conservative parameter selection gets N2V back in track but departs from its fast mapping intention.

Sentiment Classification Fig.4 shows the results of Amazon Review sentiment classification. Mem2Vec delivers impressive performance, beating all the baselines in 10 of the total 12 pairs, including CRE and DAREP. This demonstrates the advantage of the memory as a transfer medium over the word- frequency based transfer of CRE and DAREP. But CRE and DAREP are still strong baselines in the cross domain task, surpassing SCL by a large margin. N2V and SUM are built for learning representation from small data, but they don't consider the possible domain discrepancy when using pre-trained embeddings on the target small text. So they don't bring much improvements over Word2Vec and GloVe. This also reminds us that to get the few-shot word representation learning methods in practical use, domain shift should be properly addressed.

6 Conclusion

We presented an integrated representation learning scheme which gradually learns from a big corpora and quickly adapt on tiny data. It accumulates knowledge with a long-term memory to adapt the representation learning of a novel word, in the few-shot learning case. Such adaptation means to constrain the "guess" of a DSM for the novel word according to the most relevant representation learning experience, inspired by hypothesis constraint theory for fast mapping. Experiments show the proposed method learns high quality representation from both highly informative and less definitional contexts in limited size. Pre-trained embeddings with our model also lead to improvements in Named Entity Recognition and sentiment analysis.

This work is our effort towards an ideal word representation learning scheme which learns from any amount of data. In the future work, we will explore more effective memory addressing and updating approaches to boost the few-shot representation learning. We believe not all examples are equally important and worth memorizing. Learning to memorize core examples should alleviate the data-hungry of representation learning methods.

Acknowledgments

The research work descried in this paper has been supported by the National Key Research and Development Program of China under Grant No. 2017YFB1002103 and also supported by the Natural Science Foundation of China under Grant No. 61333018.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *Proceedings of international conference on learning representations (ICLR)*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics.

Danushka Bollegala, Takanori Maehara, and Ken-ichi Kawarabayashi. 2015. Unsupervised cross-domain word representation learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. 2016. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338.

Edouard Grave, Armand Joulin, and Nicolas Usunier. 2017. Improving neural language models with a continuous cache.

Aurélie Herbelot and Marco Baroni. 2017. High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 conference on empirical methods in natural language processing (EMNLP)*.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*.

Lukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. 2017. Learning to remember rare events. *ICLR*.

Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. A la carte embedding: Cheap but effective induction of semantic feature vectors. In *Proceedings of ACL, year=2018, organization=Association for Computational Linguistics*.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of bionlp shared task 2011. In *Proceedings of the BioNLP shared task 2011 workshop*, pages 1–6. Association for Computational Linguistics.

Max Kisselew, Sebastian Padó, Alexis Palmer, and Jan Šnajder. 2015. Obtaining a better understanding of distributional models of german derivational morphology. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 58–63.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Angeliki Lazaridou, Marco Marelli, and Marco Baroni. 2017. Multimodal word meaning induction from minimal exposure to natural text. *Cognitive science*, 41(S4):677–705.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1517–1526.

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Teng Long, Emmanuel Bengio, Ryan Lowe, Jackie Chi Kit Cheung, and Doina Precup. 2017. World knowledge for reading comprehension: Rare entity prediction with hierarchical lstms using external descriptions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 825–834.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *Proceedings of international conference on learning representations (ICLR)*.

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.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents.

Tsendsuren Munkhdalai and Hong Yu. 2017. Meta networks.

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.

Sampo Pyysalo and Sophia Ananiadou. 2013. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning*, pages 1842–1850.

Pablo Sprechmann, Siddhant M Jayakumar, Jack W Rae, Alexander Pritzel, Adrià Puigdomènech Badia, Benigno Uria, Oriol Vinyals, Demis Hassabis, Razvan Pascanu, and Charles Blundell. 2018. Memory-based parameter adaptation.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. 2016. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Shaonan Wang, Jiajun Zhang, Nan Lin, and Chengqing Zong. 2017. Investigating inner properties of multimodal representation and semantic compositionality with brain-based componential semantics. In *AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5964–5972.

Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2018. Learning multimodal word representation via dynamic fusion methods. pages 5973–5980.

Shaonan Wang and Chengqing Zong. 2017. Comparison study on critical components in composition model for phrase representation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(3):16.

Dirk Weissenborn. 2017. Reading twice for natural language understanding. *CoRR*, abs/1706.02596.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1219–1228. ACM.

Fei Xu and Joshua B Tenenbaum. 2007. Word learning as bayesian inference. *Psychological review*, 114(2):245.

Wei Yang, Wei Lu, and Vincent Zheng. 2017. A simple regularization-based algorithm for learning cross-domain word embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2898–2904.