

# Topic Sensitive Attention on Generic Corpora Corrects Sense Bias in Pretrained Embeddings

Vihari Piratla\*  
IIT Bombay

Sunita Sarawagi  
IIT Bombay

Soumen Chakrabarti  
IIT Bombay

## Abstract

Given a small corpus  $\mathcal{D}_T$  pertaining to a limited set of focused topics, our goal is to train embeddings that accurately capture the sense of words in the topic in spite of the limited size of  $\mathcal{D}_T$ . These embeddings may be used in various tasks involving  $\mathcal{D}_T$ . A popular strategy in limited data settings is to adapt pretrained embeddings  $\mathcal{E}$  trained on a large corpus. To correct for sense drift, fine-tuning, regularization, projection, and pivoting have been proposed recently. Among these, regularization informed by a word’s corpus frequency performed well, but we improve upon it using a new regularizer based on the stability of its cooccurrence with other words. However, a thorough comparison across ten topics, spanning three tasks, with standardized settings of hyper-parameters, reveals that even the best embedding adaptation strategies provide small gains beyond well-tuned baselines, which many earlier comparisons ignored. In a bold departure from adapting pretrained embeddings, we propose using  $\mathcal{D}_T$  to probe, attend to, and borrow fragments from any large, topic-rich source corpus (such as Wikipedia), which need not be the corpus used to pretrain embeddings. This step is made scalable and practical by suitable indexing. We reach the surprising conclusion that even limited corpus augmentation is more useful than adapting embeddings, which suggests that non-dominant sense information may be irrevocably obliterated from pretrained embeddings and cannot be salvaged by adaptation.

## 1 Introduction

Word embeddings (Mikolov et al., 2013; Pennington et al., 2014) benefit many natural language processing (NLP) tasks. Often, a group of tasks may involve a limited corpus  $\mathcal{D}_T$  pertaining to a few focused topics, e.g., discussion boards on

Physics, video games, or Unix, or a forum for discussing medical literature. Because  $\mathcal{D}_T$  may be too small to train word embeddings to sufficient quality, a prevalent practice is to harness general-purpose embeddings  $\mathcal{E}$  pretrained on a broad-coverage corpus, not tailored to the topics of interest. The pretrained embeddings are sometimes used as-is (‘pinned’). Even if  $\mathcal{E}$  is trained on a ‘universal’ corpus, considerable sense shift may exist in the meaning of polysemous words and their cooccurrences and similarities with other words. In a corpus about Unix, ‘cat’ and ‘print’ are more similar than in Wikipedia. ‘Charge’ and ‘potential’ are more related in a Physics corpus than in Wikipedia. Thus, pinning can lead to poor target task performance in case of serious sense mismatch. Another popular practice is to initialize the target embeddings to the pretrained vectors, but then “fine-tune” using  $\mathcal{D}_T$  to improve performance in the target (Mou et al., 2015; Min et al., 2017; Howard and Ruder, 2018). As we shall see, the number of epochs of fine-tuning is a sensitive knob — excessive fine-tuning might lead to “catastrophic forgetting” (Kirkpatrick et al., 2017) of useful word similarities in  $\mathcal{E}$ , and too little fine-tuning may not adapt to target sense.

Even if we are given development (‘dev’) sets for target tasks, the best balancing act between a pretrained  $\mathcal{E}$  and a topic-focused  $\mathcal{D}_T$  is far from clear. Should we fine-tune (all word vectors) in epochs and stop when dev performance deteriorates? Or should we keep some words close to their pretrained embeddings (a form of regularization) and allow others to tune more aggressively? On what properties of  $\mathcal{E}$  and  $\mathcal{D}_T$  should the regularization strength of each word depend? Our first contribution is a new measure of semantic drift of a word from  $\mathcal{E}$  to  $\mathcal{D}_T$ , which can be used to control the regularization strength. In terms of perplexity, we show that this is superior to both epoch-

\* vihari@cse.iitb.ac.in

based tuning, as well as regularization based on simple corpus frequencies of words (Yang et al., 2017). Yet another option is to learn projections to align generic embeddings to the target sense (Bollegala et al., 2015; Barnes et al., 2018; K Sarma et al., 2018), or to a shared common space (Yin and Schütze, 2016; Coates and Bollegala, 2018; Bollegala and Bao, 2018) However, in carefully controlled experiments, none of the proposed approaches to adapting pretrained embeddings consistently beats the trivial baseline of discarding them and training afresh on  $\mathcal{D}_T$ !

Our second contribution is to explore other techniques beyond adapting generic embeddings  $\mathcal{E}$ . Often, we might additionally have easy access to a broad corpus  $\mathcal{D}_S$  like Wikipedia.  $\mathcal{D}_S$  may span many diverse topics, while  $\mathcal{D}_T$  focuses on one or few, so there may be large *overall* drift from  $\mathcal{D}_S$  to  $\mathcal{D}_T$  too. However, a judicious *subset*  $\widehat{\mathcal{D}}_S \subset \mathcal{D}_S$  may exist that would be excellent for augmenting  $\mathcal{D}_T$ . The large size of  $\mathcal{D}_S$  is not a problem: we use an inverted index that we probe with documents from  $\mathcal{D}_T$  to efficiently identify  $\widehat{\mathcal{D}}_S$ . Then we apply a novel perplexity-based joint loss over  $\widehat{\mathcal{D}}_S \cup \mathcal{D}_T$  to fit adapted word embeddings. While most of recent research focus has been on designing better methods of adapting pretrained embeddings, we show that retraining with selected source *text* is significantly more accurate than the best of embeddings-only strategy, while runtime overheads are within practical limits.

An important lesson is that non-dominant sense information may be irrevocably obliterated from generic embeddings; it may not be possible to salvage this information by post-facto adaptation.

Summarizing, our contributions are:

- We propose new formulations for training topic-specific embeddings on a limited target corpus  $\mathcal{D}_T$  by (1) adapting generic pre-trained word embeddings  $\mathcal{E}$ , and/or (2) selecting from any available broad-coverage corpus  $\mathcal{D}_S$ .
- We perform a systematic comparison of our and several recent methods on three tasks spanning ten topics and offer many insights.
- Our selection of  $\widehat{\mathcal{D}}_S$  from  $\mathcal{D}_S$  and joint perplexity minimization on  $\widehat{\mathcal{D}}_S \cup \mathcal{D}_T$  perform better than pure embedding adaptation methods, at the (practical) cost of processing  $\mathcal{D}_S$ .
- We evaluate our method even with contextual embeddings. The relative performance of the adaptation alternatives remain fairly sta-

ble whether the adapted embeddings are used on their own, or concatenated with context-sensitive embeddings (Peters et al., 2018; Cer et al., 2018).

## 2 Related work and baselines

### CBOW

We review the popular CBOW model for learning unsupervised word representations (Mikolov et al., 2013). As we scan the corpus, we collect a *focus* word  $w$  and a set  $C$  of *context* words around it, with corresponding embedding vectors  $\mathbf{u}_w \in \mathbb{R}^n$  and  $\mathbf{v}_c \in \mathbb{R}^n$ , where  $c \in C$ . The two embedding matrices  $U, V$  are estimated as:

$$\max_{U, V} \sum_{(w, C) \in \mathcal{D}} \sigma(\mathbf{u}_w \cdot \mathbf{v}_C) + \sum_{\bar{w} \sim \mathcal{D}} \sigma(-\mathbf{u}_{\bar{w}} \cdot \mathbf{v}_C) \quad (1)$$

Here  $\mathbf{v}_C$  is the average of the context vectors in  $C$ .  $\bar{w}$  is a negative focus word sampled from a slightly distorted unigram distribution of  $\mathcal{D}$ . Usually downstream applications use only the embedding matrix  $U$ , with each word vector scaled to unit length. Apart from CBOW, Mikolov et al. (2013) defined the related skipgram model, and (Pennington et al., 2014) proposed the Glove model, which can also be used in our framework. We found CBOW to work better for our downstream tasks.

### Src, Tgt and Concat baselines

In the ‘Src’ option, pre-trained embeddings  $\mathbf{u}_w^S$  trained only on a large corpus are used as-is. The other extreme, called ‘Tgt’, is to train word embeddings from scratch on the limited target corpus  $\mathcal{D}_T$ . In our experiments we found that Src performs much worse than Tgt, indicating the presence of significant drift in prominent word senses. Two other simple baselines, are ‘Concat’, that concatenates the source and target trained embeddings and let the downstream task figure out their relative roles, and ‘Avg’ that following (Coates and Bollegala, 2018) takes their simple average. Another option is to let the downstream task learn to combine multiple embeddings as in (Zhang et al., 2016).

As word embeddings have gained popularity for representing text in learning models, several methods have been proposed for enriching small datasets with pre-trained embeddings.

### Adapting pre-trained embeddings

**SrcTune:** A popular method (Min et al., 2017; Wang et al., 2017; Howard and Ruder, 2018) is to use the source embeddings  $\mathbf{u}_w^S$  to initialize  $\mathbf{u}_w$

and thereafter train on  $\mathcal{D}_T$ . We call this ‘SrcTune’. Fine-tuning requires careful control of the number of epochs with which we train on  $\mathcal{D}_T$ . Excessive training can wipe out any benefit of the source because of catastrophic forgetting. Insufficient training may not incorporate target corpus senses in case of polysemous words, and adversely affect target tasks (Mou et al., 2015). The number of epochs can be controlled using perplexity on a held-out  $\mathcal{D}_T$ , or using downstream tasks. Howard and Ruder (2018) propose to fine-tune a whole language model using careful differential learning rates. However, epoch-based termination may be inadequate. Different words may need diverse trade-offs between the source and target topics, which we discuss next.

**RegFreq (frequency-based regularization):** Yang et al. (2017) proposed to train word embeddings using  $\mathcal{D}_T$ , but with a regularizer to prevent a word  $w$ ’s embedding from drifting too far from the source embedding ( $\mathbf{u}_w^S$ ). The weight of the regularizer is meant to be inversely proportional to the concept drift of  $w$  across the two corpus. Their limitation was that corpus frequency was used as a surrogate for stability; high stability was awarded to only words frequent in both corpora. As a consequence, very few words in a focused  $\mathcal{D}_T$  about Physics will benefit from a broad coverage corpus like Wikipedia. Thousands of words like *galactic*, *stars*, *motion*, *x-ray*, and *momentum* will get low stability, although their prominent sense is the same in the two corpora. We propose a better regularization scheme in this paper. Unlike us, Yang et al. (2017) did not compare with fine-tuning.

**Projection-based methods** attempt to project embeddings of one kind to another, or to a shared common space. Bollegala et al. (2014) and Barnes et al. (2018) proposed to learn a linear transformation between the source and target embeddings. Yin and Schütze (2016) transform multiple embeddings to a common ‘meta-embedding’ space. Simple averaging are also shown to be effective (Coates and Bollegala, 2018), and a recent (Bollegala and Bao, 2018) auto-encoder based meta-embedder (AEME) is the state of the art. K Sarma et al. (2018) proposed CCA to project both embeddings to a common sub-space. Some of these methods designate a subset of the overlapping words as pivots to bridge the target and source parameters in various ways (Blitzer et al., 2006;

Ziser and Reichart, 2018; Bollegala et al., 2015). Many such techniques were proposed in a cross-domain setting, and specifically for the sentiment classification task. Gains are mainly from effective transfer of sentiment representation across domains. Our challenge arises when a corpus with broad topic coverage pretrains dominant word senses quite different from those needed by tasks associated with narrower topics.

### Language models for task transfer

Complementary to the technique of adapting individual word embeddings is the design of deeper sequence models for task-to-task transfer. Cer et al. (2018); Subramanian et al. (2018) propose multi-granular transfer of sentence and word representations across tasks using Universal Sentence Encoders. ELMo (Peters et al., 2018) trains a multi-layer sequence model to build a context-sensitive representation of words in a sentence. ULMFiT (Howard and Ruder, 2018) present additional tricks such as gradual unfreezing of parameters layer-by-layer, and exponentially more aggressive fine-tuning toward output layers. Devlin et al. (2018) propose a deep bidirectional language model for generic contextual word embeddings. We show that our topic-sensitive embeddings provide additional benefit even when used with contextual embeddings.

## 3 Proposed approaches

We explore two families of methods: (1) those that have access to only pretrained embeddings (Sec 3.1), and (2) those that also have access to a source corpus with broad topic coverage (Sec 3.2).

### 3.1 RegSense: Stability-based regularization

Our first contribution is a more robust definition of stability to replace the frequency-based regularizer of RegFreq. We first train word vectors on  $\mathcal{D}_T$ , and assume the pretrained embeddings  $\mathcal{E}$  are available. Let the focus embeddings of word  $w$  in  $\mathcal{E}$  and  $\mathcal{D}_T$  be  $\mathbf{u}_w^S$  and  $\mathbf{u}_w^T$ . We overload  $\mathcal{E} \cap \mathcal{D}_T$  as words that occur in both. For each word  $w \in \mathcal{E} \cap \mathcal{D}_T$ , we compute  $N_S^{(K)}(w, \mathcal{E} \cap \mathcal{D}_T)$ , the  $K$  nearest neighbors of  $w$  with respect to the generic embeddings, i.e., with the largest values of  $\cos(\mathbf{u}_w^S, \mathbf{u}_n^S)$  from  $\mathcal{E} \cap \mathcal{D}_T$ . Here  $K$  is a suitable hyperparameter. Now we define  $\text{stability}(w) =$

$$\frac{\sum_{n \in N_S^{(K)}(w, \mathcal{E} \cap \mathcal{D}_T)} \cos(\mathbf{u}_w^T, \mathbf{u}_n^T)}{|N_S^{(K)}(w, \mathcal{E} \cap \mathcal{D}_T)|} \quad (2)$$

Intuitively, if we consider near neighbors  $n$  of  $w$  in terms of source embeddings, and most of these  $n$ 's have target embeddings very similar to the target embedding of  $w$ , then  $w$  is stable across  $\mathcal{E}$  and  $\mathcal{D}_T$ , i.e., has low semantic drift from  $\mathcal{E}$  to  $\mathcal{D}_T$ .

While many other forms of stability can achieve the same ends, ours seems to be the first formulation that goes beyond mere word frequency and employs the topological stability of near-neighbors in the embedding space. Here is why this is important. Going from a generic corpus like Wikipedia to the very topic-focused StackExchange (Physics) corpus  $\mathcal{D}_T$ , the words *x-ray*, *universe*, *kilometers*, *nucleons*, *absorbs*, *emits*, *sqrt*, *anode*, *diodes*, and *km/h* have large stability per our definition above, but low stability according to Yang et al.'s frequency method since they are (relatively) rare in source. Using their method, therefore, these words will not benefit from reliable pretrained embeddings.

Finally, the word regularization weight is:

$$R(w) = \max(0, \tanh(\lambda \text{stability}(w))). \quad (3)$$

Here  $\lambda$  is a hyperparameter.  $R(w)$  above is a replacement for the regularizer used by Yang et al. (2017). If  $R(w)$  is large, it is regularized more heavily toward its source embedding, keeping  $\mathbf{u}_w$  closer to  $\mathbf{u}_w^S$ . The modified CBOW loss is:

$$\max_{\mathbf{U}, \mathbf{V}} \sum_{\langle w, C \rangle \in \mathcal{D}} \sigma(\mathbf{u}_w \cdot \mathbf{v}_C) + \sum_{\bar{w} \sim \mathcal{D}} \sigma(-\mathbf{u}_{\bar{w}} \cdot \mathbf{v}_C) + \sum_w R(w) \|\mathbf{u}_w - \mathbf{u}_w^S\|^2 \quad (4)$$

Our  $R(w)$  performs better than Yang et al.'s.

### 3.2 Source selection and joint perplexity

To appreciate the limitations of regularization, consider words like *potential*, *charge*, *law*, *field*, *matter*, *medium*, etc. These will get small stability ( $R(w)$ ) values because their dominant senses in a universal corpus do not match with those in a Physics corpus ( $\mathcal{D}_T$ ), but  $\mathcal{D}_T$  may be too limited to wipe that dominant sense for a subset of words while preserving the meaning of stable words. However, there are plenty of high-quality broad-coverage sources like Wikipedia that includes plenty of Physics documents that could gainfully supplement  $\mathcal{D}_T$ . Therefore, we seek to include target-relevant documents from a generic source corpus  $\mathcal{D}_S$ , even if the dominant sense of a word in  $\mathcal{D}_S$  does not match that in  $\mathcal{D}_T$ . The goal is to do this without solving the harder problem of unsupervised, expensive and imperfect sense dis-

covery in  $\mathcal{D}_S$  and sense tagging of  $\mathcal{D}_T$ , and using per-sense embeddings.

The main steps of the proposed approach, SrcSel, are shown in Figure 1. Before describing the steps in detail, we note that preparing and probing a standard inverted index (Baeza-Yates and Ribeiro-Neto, 1999) are extremely fast, owing to decades of performance optimization. Also, index preparation can be amortized over multiple target tasks. (The granularity of a ‘document’ can be adjusted to the application.)

- 1: Index all source docs  $\mathcal{D}_S$  in a text retrieval engine.
- 2: Initialize a score accumulator  $a_s$  for each source doc  $s \in \mathcal{D}_S$ .
- 3: **for** each target doc  $t \in \mathcal{D}_T$  **do**
- 4:     Get source docs most similar to  $t$ .
- 5:     Augment their score accumulators.
- 6:  $\widehat{\mathcal{D}}_S \leftarrow \emptyset$
- 7: **for** each source doc  $s \in \mathcal{D}_S$  **do**
- 8:     **if**  $a_s$  is “sufficiently large” **then**
- 9:         Add  $s$  to  $\widehat{\mathcal{D}}_S$ .
- 10: Fit word embeddings to optimize a joint objective over  $\widehat{\mathcal{D}}_S \cup \mathcal{D}_T$ .

Figure 1: Main steps of SrcSel.

**Selecting source documents to retain:** Let  $s \in \mathcal{D}_S, t \in \mathcal{D}_T$  be source and target documents. Let  $\text{sim}(s, t)$  be the similarity between them, in terms of the TFIDF cosine score commonly used in Information Retrieval (Baeza-Yates and Ribeiro-Neto, 1999). The total vote of  $\mathcal{D}_T$  for  $s$  is then  $\sum_{t \in \mathcal{D}_T} \text{sim}(s, t)$ . We choose a suitable cutoff on this aggregate score, to reduce  $\mathcal{D}_S$  to  $\widehat{\mathcal{D}}_S$ , as follows. Intuitively, if we hold out a randomly sampled part of  $\mathcal{D}_T$ , our cutoff should let through a large fraction (we used 90%) of the held-out part. Once we find such a cutoff, we apply it to  $\mathcal{D}_S$  and retain the source documents whose aggregate scores exceed the cutoff. Beyond mere selection, we design a joint perplexity objective over  $\widehat{\mathcal{D}}_S \cup \mathcal{D}_T$ , with a term for the amount of trust we place in a retained source document. This limits damage from less relevant source documents that slipped through the text retrieval filter. Since the retained documents are weighted based on their relevance to the topical target corpus  $\mathcal{D}_T$ , we found it beneficial to also include a percentage (we used 10%) of randomly selected documents from  $\mathcal{D}_S$ . We refer to the method that only uses documents retained

using text retrieval filter as SrcSel:R and only randomly selected documents from  $\mathcal{D}_S$  as SrcSel:c. SrcSel uses documents both from the retrieval filter and random selection.

**Joint perplexity objective:** Similar to Eqn. (1), we will sample word and context  $\langle w, C \rangle$  from  $\mathcal{D}_T$  and  $\widehat{\mathcal{D}}_S$ . Given our limited trust in  $\widehat{\mathcal{D}}_S$ , we will give each sample from  $\widehat{\mathcal{D}}_S$  an alignment score  $Q(w, C)$ . This should be large when  $w$  is used in a context similar to contexts in  $\mathcal{D}_T$ . We judge this based on the target embedding  $\mathbf{u}_w^T$ :

$$Q(w, C) = \max \{0, \cos(\mathbf{u}_w^T, \mathbf{v}_C^T)\}. \quad (5)$$

Since  $\mathbf{u}_w$  represents the sense of the word in the target, source contexts  $C$  which are similar will get a high score. Similarity in source embeddings is not used here because our intent is to preserve the target senses. We tried other forms such as dot-product or its exponential and chose the above form because it is bounded and hence less sensitive to gross noise in inputs.

The word2vec objective (1) is enhanced to

$$\sum_{\langle w, C \rangle \in \mathcal{D}_T} \left[ \sigma(\mathbf{u}_w \cdot \mathbf{v}_C) + \sum_{\bar{w} \sim \mathcal{D}_T} \sigma(-\mathbf{u}_{\bar{w}} \cdot \mathbf{v}_C) \right] + \sum_{\langle w, C \rangle \in \widehat{\mathcal{D}}_S} Q(w, C) \left[ \sigma(\mathbf{u}_w \cdot \mathbf{v}_C) + \sum_{\bar{w} \sim \widehat{\mathcal{D}}_S} \sigma(-\mathbf{u}_{\bar{w}} \cdot \mathbf{v}_C) \right]. \quad (6)$$

The first sum is the regular word2vec loss over  $\mathcal{D}_T$ . Word  $\bar{w}$  is sampled from the vocabulary of  $\mathcal{D}_T$  as usual, according to a suitable distribution. The second sum is over the retained source documents  $\widehat{\mathcal{D}}_S$ . Note that  $Q(w, C)$  is computed using the pre-trained target embeddings and does not change during the course of training.

**SrcSel+RegSense combo:** Here we combine objective (6) with the regularization term in (4), where  $R$  uses all of  $\mathcal{E}$  as in RegSense.

## 4 Experiments

We compare the methods discussed thus far, with the goal of answering these research questions:

1. Can word-based regularization (RegFreq and RegSense) beat careful termination at epoch granularity, after initializing with source embeddings (SrcTune)?
2. How do these compare with just fusing Src and Tgt via recent meta-embedding methods like AAEME (Bollegala and Bao, 2018)<sup>1</sup>?

<sup>1</sup>We used the implementation available at: <https://github.com/CongBao/AutoencodedMetaEmbedding>

3. Does SrcSel provide sufficient and consistent gains over RegSense to justify the extra effort of processing a source corpus?
4. Do contextual embeddings obviate the need for adapting word embeddings?

We also establish that initializing with source embeddings also improves regularization methods. (Curiously, RegFreq was never combined with source initialization.)

### Topics and tasks

We compare across 15 topic-task pairs spanning 10 topics and 3 task types: an unsupervised language modeling task on five topics, a document classification task on six topics, and a duplicate question detection task on four topics. In our setting,  $\mathcal{D}_T$  covers a small subset of topics in  $\mathcal{D}_S$ , which is the 20160901<sup>2</sup> version dump of Wikipedia. Our tasks are different from GLUE-like multi-task learning (Wang et al., 2019), because our focus is on the problems created by the divergence between prominent sense-dominated generic word embeddings and their sense in narrow target topics. We do not experiment on the cross-domain sentiment classification task popular in domain adaptation papers since they benefit more from sharing sentiment-bearing words, than learning the correct sense of polysemous words, which is our focus here. All our experiments are on public datasets, and we will publicly release our experiment scripts and code.

**StackExchange topics** We pick four topics (Physics, Gaming, Android and Unix) from the CQADupStack<sup>3</sup> dataset of questions and responses. For each topic, the available response text is divided into  $\mathcal{D}_T$ , used for training/adapting embeddings, and  $\mathcal{D}_S$ , the evaluation fold used to measure perplexity. In each topic, the target corpus  $\mathcal{D}_T$  has 2000 responses totalling roughly 1 MB. We also report results with changing sizes of  $\mathcal{D}_T$ . Depending on the method we use  $\mathcal{D}_T$ ,  $\mathcal{D}_S$ , or  $\mathbf{u}^S$  to train topic-specific embeddings and evaluate them as-is on two tasks that train task-specific layers on top of these fixed embeddings. The first is an **unsupervised language modeling task** where we train a LSTM<sup>4</sup> on the adapted embed-

<sup>2</sup>The target corpora in our experiments came from datasets that were created before this time.

<sup>3</sup><http://nlp.cis.unimelb.edu.au/resources/cqadupstack/>

<sup>4</sup>[https://github.com/tensorflow/models/blob/master/tutorials/rnn/ptb/ptb\\_word\\_lm.py](https://github.com/tensorflow/models/blob/master/tutorials/rnn/ptb/ptb_word_lm.py)

Method	Physics	Gaming	Android	Unix
Tgt	121.9	185.0	142.7	159.5
Tgt(unpinned)	-0.6	-0.8	0.2	0.1

Table 1: Average reduction in perplexity, when embeddings are not pinned, on four Stackexchange topics.

dings (which are pinned) and report perplexity on  $\mathcal{D}_T$ . The second is a **Duplicate question detection task**. Available in each topic are human annotated duplicate questions (statistics in Table 10 of Appendix) which we partition across train, test and dev as 50%, 40%, 10%. For contrastive training, we add four times as much randomly chosen non-duplicate pairs. The goal is to predict duplicate/not for a question pair, for which we use word mover distance (Kusner et al., 2015, WMD) over adapted word embeddings. We found WMD more accurate than BiMPM (Wang et al., 2017). We use three splits of the target corpus, and for each resultant embedding, measure AUC on three random (train-)dev-test splits of question pairs, for a total of nine runs. For reporting AUC, WMD does not need the train fold.

**Medical domain:** This domain from the Ohsumed<sup>5</sup> dataset has abstracts on cardiovascular diseases. We sample 1.4 MB of abstracts as target corpus  $\mathcal{D}_T$ . We evaluate embeddings on two tasks: (1) unsupervised language modeling on remaining abstracts, and (2) supervised classification on 23 MeSH classes based on title. We randomly select 10,000 titles with train, test, dev split as 50%, 40%, and 10%. Following Joulin et al. (2017), we train a softmax layer on the average of adapted (and pinned) word embeddings.

**Topics from 20 newsgroup** We choose the five top-level classes in the 20 newsgroup dataset<sup>6</sup> as topics; viz.: *Computer, Recreation, Science, Politics, Religion*. The corresponding five downstream tasks are text classification over the 3–5 fine-grained classes under each top-level class. Train, test, dev splits were 50%, 40%, 10%. We average over nine splits. The `body` text is used as  $\mathcal{D}_T$  and `subject` text is used for classification.

Pretrained embeddings  $\mathcal{E}$  are trained on Wikipedia using the default settings of word2vec’s CBOW model. All our data splits are made publicly available at [https://github.com/vihari/we\\_adapt\\_datasets](https://github.com/vihari/we_adapt_datasets).

<sup>5</sup><https://www.mat.unical.it/OlexSuite/Datasets/SampleDataSets-about.htm>

<sup>6</sup><http://qwone.com/~jason/20Newsgroups/>

#### 4.1 Effect of fine-tuning embeddings on the target task

We chose to pin embeddings in all our experiments, once adapted to the target corpus, namely the document classification task on medical and 20 newsgroup topics and language model task on five different topics. This is because we did not see any improvements when we unpin the input embeddings. We summarize in Table 1 the results when the embeddings are not pinned on language model task on the four StackExchange topics.

#### 4.2 Epochs vs. regularization results

In Figure 2 we show perplexity and AUC against training epochs. Here we focus on four methods: Tgt, SrcTune, RegFreq, and RegSense. First note that Tgt continues to improve on both perplexity and AUC metrics beyond five epochs (the default in word2vec code<sup>7</sup> and left unchanged in RegFreq<sup>8</sup> (Yang et al., 2017)). In contrast, SrcTune, RegSense, and RegFreq are much better than Tgt at five epochs, saturating quickly. With respect to perplexity, SrcTune starts getting worse around 20 iterations and becomes identical to Tgt, showing catastrophic forgetting. Regularizers in RegFreq and RegSense are able to reduce such forgetting, with RegSense being more effective than RegFreq. These experiments show that any comparison that chooses a fixed number of training epochs across all methods is likely to be unfair. Henceforth we will use a validation set for the stopping criteria. While this is standard practice for supervised tasks, most word embedding code we downloaded ran for a fixed number of epochs, making comparisons unreliable. We conclude that validation-based stopping is critical for fair evaluation.

We next compare SrcTune, RegFreq, and RegSense on the three tasks: perplexity in Table 2, duplicate detection in Table 3, and classification in Table 4. All three methods are better than baselines Src and Concat, which are much worse than Tgt indicating the presence of significant concept drift. Yang et al. (2017) provided no comparison between RegFreq (their method) and SrcTune; we find the latter slightly better. On the supervised tasks, RegFreq is often worse than Tgt provided Tgt is allowed to train for enough epochs.

<sup>7</sup><https://code.google.com/archive/p/word2vec/>

<sup>8</sup>[https://github.com/Victor0118/cross\\_domain\\_embedding/](https://github.com/Victor0118/cross_domain_embedding/)

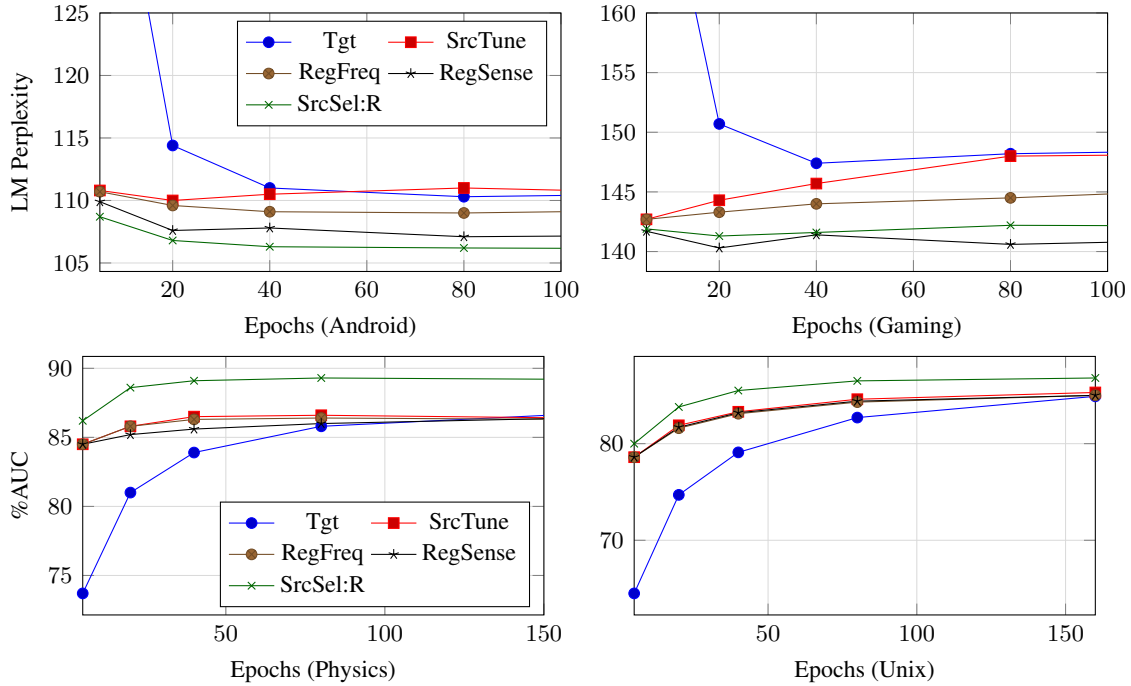


Figure 2: Language model perplexity (top row) and AUC on duplicate question detection (bottom row).

Method	Physics	Gaming	Android	Unix	Med
Tgt	121.9	185.0	142.7	159.5	158.9
SrcTune	2.3	6.8	1.1	3.1	5.5
RegFreq	2.1	7.1	1.8	3.4	6.8
RegSense	5.0	13.8	6.7	9.7	14.6
SrcSel	5.8	11.7	5.9	6.4	8.6
SrcSel+RegSense	6.2	12.5	7.9	9.3	10.5

Table 2: Average reduction in language model perplexity over Tgt on five topics.  $\pm$  standard deviation are shown in Table 11 in the Appendix

If the same number of epochs are used to train the two methods, one can reach the misleading conclusion that Tgt is worse. RegSense is better than SrcTune and RegFreq particularly with respect to perplexity, and rare class classification (Table 4). We conclude that a well-designed word stability-based regularizer can improve upon epoch-based fine-tuning.

**Impact of source initialization** Table 5 compares Tgt and RegFreq with two initializers: (1) random as proposed by Yang et al. (2017), and (2) with source embeddings. RegFreq after source initialization is better in almost all cases. SrcSel and RegSense also improve with source initialization, but to a smaller extent. (More detailed numbers are in Table 14 of Appendix.) We conclude that initializing with pretrained embeddings is helpful even with regularizers.

	Physics	Gaming	Android	Unix
Tgt	86.7	82.6	86.8	85.4
Src	-2.3 $\pm$ 0.5	0.8 $\pm$ 0.5	-3.7 $\pm$ 0.5	-7.1 $\pm$ 0.3
Concat	-1.1 $\pm$ 0.5	1.4 $\pm$ 0.3	-2.1 $\pm$ 0.3	-4.5 $\pm$ 0.4
AAEME	1.2 $\pm$ 0.2	4.6 $\pm$ 0.0	-0.3 $\pm$ 0.2	0.0 $\pm$ 0.2
SrcTune	-0.3 $\pm$ 0.3	1.9 $\pm$ 0.2	0.6 $\pm$ 0.2	-0.0 $\pm$ 0.2
RegFreq	-0.4 $\pm$ 0.2	2.4 $\pm$ 0.2	-0.5 $\pm$ 0.5	-0.5 $\pm$ 0.2
RegSense	-0.4 $\pm$ 0.5	2.2 $\pm$ 0.1	-0.5 $\pm$ 0.5	-0.5 $\pm$ 0.4
SrcSel	3.6 $\pm$ 0.2	3.0 $\pm$ 0.2	0.8 $\pm$ 0.3	2.1 $\pm$ 0.2
SrcSel+RegSense	3.6 $\pm$ 0.2	3.1 $\pm$ 0.5	0.8 $\pm$ 0.3	2.1 $\pm$ 0.2

Table 3: AUC gains over Tgt ( $\pm$  standard deviation of difference) on duplicate question detection task on various target topics. AAEME is the auto-encoder meta-embedding of Bollegala and Bao (2018).

**Comparison with Meta-embeddings** In Tables 3 and 4 we show results with the most recent meta-embedding method AAEME. AAEME provides gains over Tgt in only two out of six cases<sup>9</sup>.

### 4.3 Performance of SrcSel

We next focus on the performance of SrcSel on all three tasks: perplexity in Table 2, duplicate detection in Table 3, and classification in Table 4. SrcSel is always among the best two methods for perplexity. In supervised tasks, SrcSel is

<sup>9</sup>On the topic classification datasets in Table 4, AAEME and its variant DAEME were worse than Src. We used the dev set to select the better of Src and their best method.

Method	Ohsumed			20NG Avg 5 topics
	Micro	Macro	Rare	
Tgt	26.3	14.7	3.0	88.9
Src	-1.0±0.9	0.±0.5	0.±0.1	-3.9±1.2
AAEME	-1.0±0.9	0.±0.5	0.±0.1	-3.9±1.2
SrcTune	1.7±1.0	1.8±1.7	1.5±2.0	0.0±1.6
RegFreq	0.6±0.5	1.8±2.3	3.7±4.7	-
RegSense	1.4±0.5	2.5±1.2	4.0±1.8	0.4±1.3
SrcSel	2.0±0.9	2.6±1.5	1.1±1.4	0.5±1.5
SrcSel	2.3±0.7	3.4±1.3	4.3±1.2	0.5±1.5
+RegSense				

Table 4: Average accuracy gains over Tgt ( $\pm$  std-dev) on Ohsumed and 20NG datasets. We show macro and rare class accuracy gains for Ohsumed because of its class population skew. Per-topic 20NG gains are in Table 15 in Appendix.

	Physics	Gaming	Android	Unix
RegFreq’s reduction in Perplexity over Tgt				
Original	1.1±1.1	1.5±1.2	0.9±0.1	0.7±0.8
+SrcInit	2.1±0.9	5.7±0.8	1.1±0.5	2.1±0.8
RegFreq’s gain in AUC over Tgt				
Original	-1.2±0.4	0.1±0.1	-0.2±0.1	-0.4±0.1
+SrcInit	-0.4±0.2	2.4±0.2	-0.5±0.5	-0.5±0.2

Table 5: Effect of initializing with source embeddings. We show mean gains over Tgt over 9 runs ( $\pm$  std-dev).

the only method that provides significant gains for all topics: AUC for duplicate detection increases by 2.4%, and classification accuracy increases by 1.4% on average. SrcSel+RegSense performs even better than SrcSel on all three tasks particularly on rare words. An ablation study on other variants of SrcSel appear in the Appendix.

**Word-pair similarity improvements:** In Table 6, we show normalized<sup>10</sup> cosine similarity of word pairs pertaining to the Physics and Unix topics. Observe how word pairs like (*nice, kill*), (*vim, emacs*) in Unix and (*current, electron*), (*lie, group*) in Physics are brought closer together as a result of importing the larger unix/physics subset from  $\mathcal{D}_S$ . In each of these pairs, words (e.g. *nice, vim, lie, current*) have a different prominent sense in the source (Wikipedia). Hence, methods like SrcTune, and RegSense cannot help. In contrast, word pairs like (*cost, require*), (*x-ray, x-rays*) whose sense is the same in the two corpus benefit significantly from the source across all methods.

<sup>10</sup>We sample a set  $S$  of 20 words based on their frequency. Normalized similarity between  $a$  and  $b$  is  $\frac{\cos(a,b)}{\sum_{w \in (S \cup b)} \cos(a,w)}$ . Set  $S$  is fixed across methods.

Pair	Tgt	Src Tune	Reg Freq	Reg Sense	Src Sel
Unix topic					
nice, kill	4.6	4.5	4.4	4.4	5.2
vim, emacs	5.7	5.8	5.7	5.8	6.4
print, cat	5.0	4.9	4.9	5.0	5.4
kill, job	5.2	5.1	5.2	5.3	5.8
make, install	5.1	5.1	5.3	5.7	5.8
character, unicode	4.9	5.1	4.7	4.6	5.8
Physics topic					
lie, group	5.2	5.0	4.4	5.1	5.8
current, electron	5.3	5.3	4.7	5.3	5.7
potential, kinetic	5.8	5.8	4.5	5.9	6.1
rotated, spinning	5.0	5.7	6.0	5.1	5.6
x-ray, x-rays	5.3	7.0	6.1	5.5	6.4
require, cost	4.9	6.2	5.2	5.1	5.3
cool, cooling	5.6	6.0	6.4	5.7	5.7

Table 6: Example word pairs and their normalized similarity across different methods of training embeddings.

**Running time:** SrcSel is five times slower than RegFreq, which is still eminently practical.  $\widehat{\mathcal{D}}_S$  was within  $3\times$  the size of  $\mathcal{D}_T$  in all domains. If  $\mathcal{D}_S$  is available, SrcSel is a practical and significantly more accurate option than adapting pretrained source embeddings. SrcSel+RegSense complements SrcSel on rare words, improves perplexity, and is never worse than SrcSel.

	Physic	Game	Andrd	Unix	Med(Rare)
Tgt	89.7	88.4	89.4	89.2	9.4
SrcTune	-0.2	0.6	-0.4	-0.2	-2.1
SrcSel	1.9	0.5	0.0	-0.2	1.1

Table 7: Performance with a larger target corpus size of 10MB on the four deduplication tasks (AUC score) and one classification task (Accuracy on rare class). Details in Table 16 of Appendix.

**Effect of target corpus size** The problem of importing source embeddings is motivated only when target data is limited. When we increase target corpus 6-fold, the gains of SrcSel and SrcTune over Tgt was insignificant in most cases. However, infrequent classes continued to benefit from the source as shown in Table 7.

#### 4.4 Contextual embeddings

We explore if contextual word embeddings obviate the need for adapting source embeddings, in the ELMo (Peters et al., 2018) setting, a contextualized word representation model, pre-trained on a 5.5B token corpus<sup>11</sup>. We compare ELMo’s

<sup>11</sup><https://allennlp.org/elmo>



	Physic	Game	Andrd	Unix	Med
Tgt	86.7	82.6	86.8	85.4	26.3
ELMo	-1.0	4.5	-1.5	-2.3	3.2
+Tgt	-0.8	3.8	0.5	0.0	4.1
+SrcTune	-0.5	3.0	0.3	0.2	3.5
+SrcSel	2.6	4.1	1.1	1.5	4.6

Table 8: Gains over Tgt with contextual embeddings on duplicate detection (columns 2–5) and classification (column 6). (Std-dev in Table 17 of Appendix.)

contextual embeddings as-is, and also after concatenating them with each of Tgt, SrcTune, and SrcSel embeddings in Table 8. First, ELMo+Tgt is better than Tgt and ELMo individually. This shows that contextual embeddings are useful but they do not eliminate the need for topic-sensitive embeddings. Second, ELMo+SrcSel is better than ELMo+Tgt. Although SrcSel is trained on data that is a strict subset of ELMo, it is still instrumental in giving gains since that subset is aligned better with the target sense of words. We conclude that topic-adapted embeddings can be useful, even with ELMo-style contextual embeddings.

Recently, BERT (Devlin et al., 2018) has garnered a lot of interest for beating contemporary contextual embeddings on all the GLUE tasks. We evaluate BERT on question duplicate question detection task on the four StackExchange topics. We use pre-trained BERT-base, a smaller 12-layer transformer network, for our experiments. We train a classification layer on the final pooled representation of the sentence pair given by BERT to obtain the binary label of whether they are duplicates. This is unlike the earlier setup where we used EMD on the fixed embeddings.

To evaluate the utility of a relevant topic focused corpus, we fine-tune the pre-trained checkpoint either on  $\mathcal{D}_T$  (SrcTune) or on  $\mathcal{D}_T \cup \widehat{\mathcal{D}}_S$  (SrcSel:R) using BERT’s masked language model loss. The classifier is then initialized with the fine-tuned checkpoint. Since fine-tuning is sensitive to the number of update steps, we tune the number of training steps using performance on a held-out dev set. F1 scores corresponding to different initializing checkpoints are shown in table 9. It is clear that pre-training the contextual embeddings on relevant target corpus helps in the downstream classification task. However, the gains of SrcSel:R over Tgt is not clear. This could be due to incomplete or noisy sentences in  $\widehat{\mathcal{D}}_S$ . There is need for more experimentation and research to understand the limited gains of SrcSel:R over SrcTune in the case of

Method	Physics	Gaming	Android	Unix
BERT	87.5	85.3	87.4	82.7
SrcTune	<b>88.0</b>	<b>89.2</b>	88.5	83.5
SrcSel:R	87.9	88.4	<b>88.6</b>	<b>85.1</b>

Table 9: F1 scores on question de-duplication task using BERT-base and when fine-tuned on Tgt only ( $\mathcal{D}_T$ ) and Tgt and selected source ( $\mathcal{D}_T \cup \widehat{\mathcal{D}}_S$ )

BERT. We leave this for future work.

## 5 Conclusion

We introduced one regularization and one source-selection method for adapting word embeddings from a partly useful source corpus to a target topic. They work better than recent embedding transfer methods, and give benefits even with contextual embeddings. It may be of interest to extend these techniques to embed knowledge graph elements.

**Acknowledgment:** Partly supported by an IBM AI Horizon grant. We thank all the anonymous reviewers for their constructive feedback.

## References

- Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Projecting embeddings for domain adaptation: Joint modeling of sentiment analysis in diverse domains. In *COLING*.
- 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 and Cong Bao. 2018. Learning word meta-embeddings by autoencoding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1650–1661.
- 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 and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, *ACL*.
- Danushka Bollegala, David J. Weir, and John A. Carroll. 2014. Learning to predict distributions of words across domains. In *ACL*.

- Daniel Cer et al. 2018. [Universal sentence encoder](#). *CoRR*, abs/1803.11175.
- Joshua Coates and Danushka Bollegala. 2018. Frustratingly easy meta-embedding - computing meta-embeddings by averaging source word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 194–198.
- 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*.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.
- Prathusha K Sarma, Yingyu Liang, and Bill Sethares. 2018. Domain adapted word embeddings for improved sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.
- Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). In *NIPS Conference*, pages 3111–3119.
- Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. *arXiv preprint arXiv:1702.02171*.
- Lili Mou, Hao Peng, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2015. Discriminative neural sentence modeling by tree-based convolution. In *EMNLP*.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. [GloVe: Global vectors for word representation](#). In *EMNLP Conference*, volume 14, pages 1532–1543.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). In *ACL Conference*, volume 1, pages 2227–2237.
- Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J. Pal. 2018. [Learning general purpose distributed sentence representations via large scale multi-task learning](#). *CoRR*, abs/1804.00079.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [Glue: A multi-task benchmark and analysis platform for natural language understanding](#). In *ICLR*. ArXiv 1804.07461.
- Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*.
- 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.
- Wenpeng Yin and Hinrich Schütze. 2016. Learning word meta-embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1351–1360.
- Ye Zhang, Stephen Roller, and Byron C. Wallace. 2016. Mgn-cnn: A simple approach to exploiting multiple word embeddings for sentence classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Yftah Ziser and Roi Reichart. 2018. Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.