Cross-Topic Distributional Semantic Representations Via Unsupervised Mappings

Eleftheria Briakou[1,2]*, **Nikos Athanasiou**[2], **Alexandros Potamianos**[2,3]

[1]University of Maryland, College Park, MD
[2]School of ECE, National Technical University of Athens, Athens, Greece
[3]Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, USA

ebriakou@cs.umd.edu, athn.nik@gmail.com, potam@central.ntua.gr

Abstract

In traditional Distributional Semantic Models (DSMs) the multiple senses of a polysemous word are conflated into a single vector space representation. In this work, we propose a DSM that learns multiple distributional representations of a word based on different topics. First, a separate DSM is trained for each topic and then each of the topic-based DSMs is aligned to a common vector space. Our unsupervised mapping approach is motivated by the hypothesis that words preserving their relative distances in different topic semantic sub-spaces constitute robust *semantic anchors* that define the mappings between them. Aligned cross-topic representations achieve state-of-the-art results for the task of contextual word similarity. Furthermore, evaluation on NLP downstream tasks shows that multiple topic-based embeddings outperform single-prototype models.

1 Introduction

Word-level representation learning algorithms adopt the *distributional hypothesis* (Harris, 1954), presuming a correlation between the distributional and the semantic relationships of words. Typically, these models encode the contextual information of words into dense feature vectors—often referred to as *embeddings*—of a k-dimensional space, thus creating a Vector Space Model (VSM) of lexical semantics. Such embeddings have been successfully applied to various natural language processing applications, including information retrieval (Manning et al., 2008), sentiment analysis (Tai et al., 2015), and machine translation (Amiri et al., 2016; Sharaf et al., 2017).

Despite their popularity, traditional DSMs rely solely on models where each word is uniquely represented by one point in the vector space. From a linguistic perspective, these models cannot capture the distinct meanings of polysemous words (e.g., *bank* or *cancer*), resulting in conflated word representations of diverse contextual semantics.

To alleviate this problem, DSMs with multiple representations per word have been proposed in the literature, based on clustering local contexts of individual words (Reisinger and Mooney, 2010; Tian et al., 2014; Neelakantan et al., 2014). An alternative way to train multiple representation DSMs is to utilize semantic lexical resources (Rothe and Schütze, 2015; Pilehvar and Collier, 2016). Christopoulou et al. (2018), based on the assumption that typically words appear with a specific sense in each topic, proposed a topic-based semantic mixture model that exploits a combination of similarities estimated on topic-based DSMs for the computation of semantic similarity between words. Their model performs well for a variety of semantic similarity tasks; however, it lacks a unified representation of multiple senses in a common semantic space. The problem of defining transformations between embeddings—trained independently under different corpora—has been previously examined in various works, such as machine translation (Mikolov et al., 2013b; Xing et al., 2015; Artetxe et al., 2016), induction of historical embeddings (Hamilton et al., 2016) and lexical resources enrichment (Prokhorov et al., 2017).

Following this line of research, we induce the creation of multiple cross-topic word embeddings by projecting the semantic representations of topic-based DSMs to a unified semantic space. We investigate different ways to perform the mappings from the topic sub-spaces to the unified semantic space, and propose a completely unsupervised approach to extract *semantic anchors* that define those mappings. Furthermore, we claim that polysemous words change their meaning in different topic domains; this is reflected in rela-

[*] The research was performed when the author was an undergraduate researcher at School of ECE, NTUA in Athens, Greece.

tive shifts of their distributional representations in different topic-based DSMs. On the other hand, semantic anchors should have consistent semantic relationships regardless of the domain they reside in. Hence, their distributions of similarity values should also be similar across different domains. Finally, we apply a smoothing technique to each word's set of topic embeddings, resulting in representations with fine-grained semantics.

To our knowledge, this is the first time that mappings between semantic spaces are applied to the problem of learning multiple embeddings for polysemous words. Our multi-topic word representations are evaluated on the contextual semantic similarity task and yield state-of-the-art performance compared to other unsupervised multi-prototype word embedding approaches. We further perform experiments on two NLP downstream tasks: text classification and paraphrase identification and demonstrate that our learned word representations consistently provide higher performance than single-prototype word embedding models. The code of the present work is publicly available[1].

2 Related Work

Methods that assign multiple distributed representations per word can be grouped into two broad categories.[2] Unsupervised methods induce multiple word representations without leveraging semantic lexical resources. Reisinger and Mooney (2010) were the first to create a multi-prototype DSM with a fixed number of vectors assigned to each word. In their model, the centroids of context-dependent clusters were used to create a set of "sense-specific" vectors for each target word. Based on similar clustering approaches, follow-up works introduced neural network architectures that incorporated both local and global context in a joint training objective (Huang et al., 2012), as well as methods that jointly performed word sense clustering and embedding learning as in Neelakantan et al. (2014); Li and Jurafsky (2015). A probabilistic framework was introduced by Tian et al. (2014), where the Skip-Gram model of Word2Vec was modified to learn multiple embedding vectors. Furthermore, latent topics were integrated into the Skip-Gram model, resulting in topical word embeddings which modeled the semantics of a word under different contexts (Liu et al., 2015b,a; Nguyen et al., 2017). Another topic-related embedding creation approach was proposed in Christopoulou et al. (2018) where a mixture of topic-based semantic models was extracted by topical adaptation of in-domain corpora. Other approaches used autoencoders (Amiri et al., 2016), convolutional neural networks designed to produce context representations that reflected the order of words in a context (Zheng et al., 2017) and reinforcement learning (Lee and Chen, 2017; Guo et al., 2018).

Supervised approaches, based on prior knowledge acquired by sense inventories (e.g., WordNet) along with word sense disambiguation algorithms, were also introduced for sense-specific representations extraction (Chen et al., 2014; Iacobacci et al., 2015). In other works, pre-trained word embeddings have been extended to embeddings of lexemes and synsets (Rothe and Schütze, 2015) or were de-conflated into their constituent sense representations (Pilehvar and Collier, 2016) by exploiting semantic lexical resources.

3 Unified Multi-Topic DSM (UTDSM)

Our system follows a four-step approach:

1. **Global Distributional Semantic Model.** Given a large collection of text data we train a DSM that encodes the contextual semantics of each word into a single representation, also referred to as Global-DSM.

2. **Topic-based Distributional Semantic Models.** Next, a topic model is trained using the same corpus. The topic model splits the corpus into K (possibly overlapping) sub-corpora. A DSM is then trained from each sub-corpus resulting in K topic-based DSMs (TDSMs). The topical adaptation of the semantic space takes into account the contextual variations a word exhibits under different thematic domains and therefore leads to the creation of "topic-specific" vectors (topic embeddings).

3. **Mappings of topic embeddings.** Next, we map the vector space of each topic-based DSM to the shared space of the Global-DSM, using a list of anchor words selected through

[1] https://github.com/Elbria/utdsm_naacl2018

[2] We limit our discussion to related works that use monolingual DSMs and corpora.

an unsupervised self-learning scheme. In the unified semantic space each word is represented by a set of topic embeddings that were previously isolated in distinct vector spaces, thus creating a Unified multi-Topic DSM (UTDSM).

4. **Smoothing of topic embeddings.** As an optional step, we employ a smoothing approach in order to cluster a word's topic embeddings into N Gaussian distributions via a Gaussian Mixture Model (GMM). This step lessens the noise introduced to our system through the semantic mappings and sparse training data.

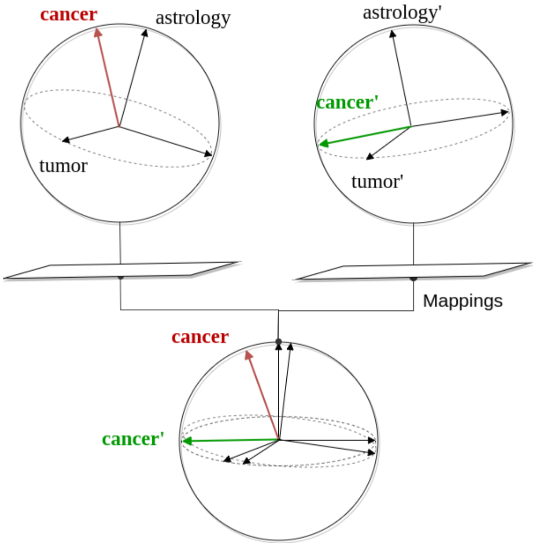


Figure 1: Simplified depiction summarizing the intuition behind the alignment process of topic embeddings. In the unified vector space, the polysemous word *cancer* is represented by two topic vectors that capture different semantic properties of the word under a zodiacal and a medical topic. Words *astrology* and *tumor* are examples of *semantic anchors* that define the mappings.

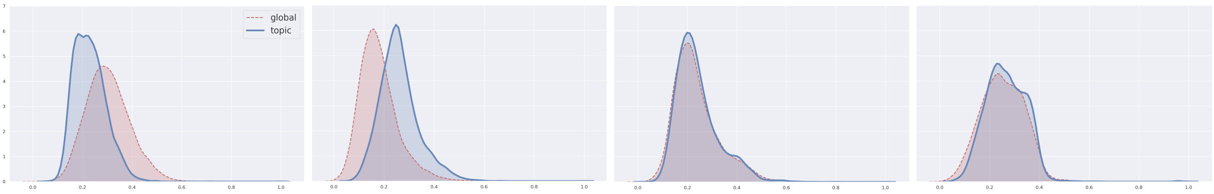
3.1 Topic-Based Distributional Semantic Models

The first step towards the thematic adaptation of the semantic space is the induction of in-domain corpora, using the Latent Dirichlet Algorithm (LDA) (Blei et al., 2003). LDA is a generative probabilistic model of a corpus. Its core idea is that documents are represented as random mixtures over topics; where each topic is defined as a probability distribution over a collection of words. Given as input a corpus of documents, LDA trains a topic model and creates a distribution of words for each topic in the corpus. Using the trained LDA model, we infer a topic distribution for each sentence in the corpus. Afterward, following a soft clustering scheme each sentence is included in a topic-specific corpus when the posterior probability for the corresponding topic exceeds a predefined threshold. The resulting topic sub-corpora are then used to train topic-based DSMs. Any of the DSM training algorithms proposed in the literature can be used for this purpose; in this paper, we opt for the Word2Vec model (Mikolov et al., 2013a).

3.2 Mappings Of Topic Embeddings

The intrinsic non-determinism of the Word2Vec algorithm leads to the creation of continuous vector spaces that are not naturally aligned to a unified semantic reference space, precluding the comparison between words of different thematic domains. To circumvent this limitation, we need to map the word representations of TDSMs to a shared vector space. In particular, we hypothesize that TDSMs capture meaningful variations in usage of polysemous words, while the relative semantic distance between monosemous words is preserved. This hypothesis motivated us to think of monosemous words as *anchors* between semantic spaces, as illustrated in Figure 1. One way to retrieve the list of anchors is to extract monosemous words from lexical resources such as WordNet (Prokhorov et al., 2017). However, this method is restricted to languages where such lexical resources exist and depends on the lexical coverage and quality of such resources.

To overcome the above limitations, we propose a fully unsupervised method for semantic anchor induction. Although the embeddings of the topic and global semantic vector spaces are not aligned, their corresponding similarity matrices (once normalized) are. Based on this observation, we compute the similarity between a given word and every other word in the vocabulary (similarity distribution) for the different topic and global spaces. Then, we assume that good semantic anchors should have similar similarity distributions across the topic-specific and the global space, as illustrated in Figure 2. Artetxe et al. (2018) was based on a similar observation to align vector semantic spaces in bilingual machine translation context.



view crater professor october

Figure 2: Similarity distributions of four different words (corresponding to the smoothed density estimates of the similarity matrices) in topic domain space as defined in Equation 1 and global space s_g^i. Selected anchors ("professor" and "october") have more similar distributions in the global and topic spaces, when compared to unselected ones ("view" and "crater"). We observe that the selected anchors are less ambiguous, while the not selected ones are expected to have diverse contextual semantics.

Let V be the intersection of the Global-DSM and the K TDSMs vocabularies and d the embedding dimension. We then define $X_k \in \mathbb{R}^{|V| \times d}$ as the embedding matrix of the k-th TDSM, and $Y \in \mathbb{R}^{|V| \times d}$ as the embedding matrix of the global DSM, where the i-th row of each matrix corresponds to the unit normalized representation of a word in V. Then, we define $S_k = X_k X_k^T$, $S_g = YY^T \in \mathbb{R}^{|V| \times |V|}$ to be the similarity distribution matrices for the k-th TDSM and the global-DSM, respectively. Then our objective is to extract a list of semantic anchors A that minimizes the Euclidean distance between the two different similarity distributions. Specifically, for every word i we calculate the average semantic distribution across all topics:

$$< s_k^i >_k = \frac{1}{K} \sum_{k=1}^{K} s_k^i \tag{1}$$

$$\| < s_k^i >_k \; - \; s_g^i \, \|_2 \, , \quad \forall \, i = 1, \ldots, |V| \tag{2}$$

where s_g^i, s_k^i is the i-th row of the S_g and S_k similarity matrix, respectively, representing the similarity distribution between word i and every other word in the vocabulary V. We then choose $|A|$ anchors as the words with the smallest values according to criterion 2. Furthermore, we assume that there exists an orthogonal transformation matrix between the topic embeddings of the extracted semantic anchors of each TDSM (source space) and the corresponding representations of the global-DSM (target space). The orthogonality constraint on the transformation matrix is widely adopted by the literature for various semantic space alignment tasks (Xing et al., 2015; Artetxe et al., 2016; Hamilton et al.). Assume $\alpha_k^j \in \mathbb{R}^d$ is the vector representation of the j-th *anchor* word in the source space and $\alpha_g^j \in \mathbb{R}^d$ is its corresponding vector representation in the target space. The transformation matrix $M_k \in \mathbb{R}^{d \times d}$ that projects the first space to the latter is learned via solving the following constraint optimization problem:[3]

$$\min_{M_k} \sum_{j=1}^{|A|} \| M_k \alpha_k^j - \alpha_g^j \|_2^2, \;\; \text{s.t.} \;\; M_k M_k^T = \mathbb{I} \tag{3}$$

The induction of multiple topic embeddings in the unified vector space is achieved via applying Equation 3 to each TDSM. Specifically, given a word and its k-th topic distributed representation $x_k \in \mathbb{R}^d$, we compute its projected representation $x_k' \in \mathbb{R}^d$ as follows:

$$x_k' = M_k x_k \tag{4}$$

3.3 Smoothing Of Topic Embeddings

Starting from the set of aligned topic embeddings $\{x_k'\}_{k=1}^K$ for each word, we learn a Gaussian Mixture Model with N components, where closely positioned topic embeddings are assigned to the same component. This step operates as an implicit way of segmenting the space of topic embeddings for each word in order to capture more useful hyper-topics—union of topics—which better represent their different meanings. We suggest that each Gaussian distribution forms a semantically coherent unit that corresponds to closely related semantics of the target word. Subsequently, the mean vector of each Gaussian distribution is used as a representative vector of each component, leading to a new set of *smoothed* topic embeddings $\{x_n^*\}_{n=1}^N$ for each word, where $x_n^* \in \mathbb{R}^d$.

[3] This problem is known as the orthogonal Procrustes problem and it has a closed form solution as proposed in (Schönemann, 1966).

4 Experimental Setup

4.1 DSM Settings

As our initial corpus we used the English Wikipedia, containing 8.5 million articles (Turney, 2012). During the training of the topic model, we used the articles found in the Wikipedia corpus and employed the Gensim implementation of LDA (Rubenstein and Goodenough, 1965) setting the number of topics K to 50. Using a threshold of 0.1, we followed a soft-clustering approach, to bootstrap the creation of topic sub-corpora, using our trained topic model. Finally, we used Gensim's implementation of Word2Vec and Continuous Bag-of-Words method to train both the global-DSM and the TDSMs. The context window parameter of Word2Vec is set to 5, while the dimensionality d of all the constructed DSMs is equal to 300 or 500.[4]

4.2 Semantic Anchors

The number of *semantic anchors* that determine the mappings between our source and target spaces is set to $|A| = 5\,000$ [5] according to our unsupervised approach (criterion 2). Those are selected from the common set of words that are represented in all semantic spaces with $|V| \sim 12\,000$.

As a second experiment, we randomly sample $|A|$ words from the vocabulary of each TDSM to define its transformation matrix. We repeat this experiment 10 times, every time sampling a different list from the corresponding vocabulary and report average performance results.

4.3 Gaussian Mixture Model

To apply the smoothing technique on the set of a word's topic embeddings we use the Scikit-learn implementation of Gaussian Mixture Model clustering algorithm (Pedregosa et al., 2011). We initialize the mean vector of each component using k-means algorithm and the parameters of the model are estimated using Expectation-Maximization (EM) algorithm.

4.4 Contextual Semantic Similarity

To estimate the semantic similarity between a pair of words provided in sentential context, we use the standard evaluation Stanford Contextual Word Similarity (SCWS) (Huang et al., 2012) dataset which consists of $2\,003$ word-pairs with assigned semantic similarity scores computed as the average estimations of several human annotators. Following the evaluation guidelines proposed in literature, we employ the $\mathrm{AvgSimC}$ and $\mathrm{MaxSimC}$ contextual metrics, firstly discussed in Reisinger and Mooney (2010). In particular, given the word-pair (w, w'), and their provided contexts (c, c') we define:

$$\mathrm{AvgSimC}(w, w') = \\ \frac{1}{K^2} \sum_{j=1}^{K} \sum_{k=1}^{K} \mathrm{p}(j|w, c)\mathrm{p}(k|w', c')\mathrm{d}(x'_j(w), x'_k(w')), \tag{5}$$

$$\mathrm{MaxSimC}(w, w') = \mathrm{d}(\hat{x}'(w), \hat{x}'(w')), \tag{6}$$

Following the notation used in 3.2, K is the number of topics returned by the trained LDA model, x'_j is the word embedding trained on the sub-corpus corresponding to the j-th topic after being projected to the unified vector space, $\mathrm{p}(j|w, c)$ denotes the posterior probability of topic j returned by LDA given as input the context c of word w, d denotes the cosine similarity between the two input representations and finally $\hat{x}'(w) = u_{\mathrm{argmax}_{1 \leq j \leq K} \mathrm{p}(j|w,c)}(w)$ is the vector representation of word w that corresponds to the topic with the maximum posterior for c. Intuitively, a higher score in $\mathrm{MaxSimC}$ indicates the existence of more robust multi-topic word representations. On the other hand, $\mathrm{AvgSimC}$ provides a topic-based smoothed result across different embeddings.

4.5 Downstream NLP Tasks

Besides the standard evaluation benchmark of contextual word similarity, we also investigate the effectiveness of our mapped cross-topic embeddings on document and sentence level downstream NLP tasks: text classification and paraphrase identification. We report weighted-averaging precision, recall, F1-measure and accuracy performance metrics.

Text classification. We used the 20NewsGroup[6] dataset, which consists of about $20\,000$ documents. Our goal is to classify each document into one of the 20 different newsgroups based on its content.

[4] Any parameter not mentioned is set to default values of the corresponding implementations (e.g., Word2Vec, Gensim LDA).

[5] We have experimented with different values of anchors from $\{1\,000, 2\,000, 3\,000, 4\,000, 5\,000\}$ and report results for the best setup.

[6] http://qwone.com/ jason/20Newsgroups/

Paraphrase Identification. For this task we aimed at identifying whether two given sentences can be considered paraphrases or not, using the Microsoft Paraphrase dataset (Dolan et al., 2004).

Document and Sentence level representations. Given a document or a sentence D, where w_d corresponds to the d-th word in D, we extract its feature representation using three different ways:

$$\mathrm{AvgC_D} = \frac{1}{|D|}\sum_{d=1}^{|D|}\sum_{k=1}^{K}\mathrm{p}(k|D)x'_k(w_d), \tag{7}$$

$$\mathrm{Avg_D} = \frac{1}{|D|}\sum_{d=1}^{|D|}\sum_{k=1}^{K}\frac{1}{K}x'_k(w_d), \tag{8}$$

$$\begin{aligned}\mathrm{MaxC_D} &= \frac{1}{|D|}\sum_{w=1}^{|D|}x'_m(w_d)\\ \text{s.t.}\quad m &= \underset{k=1,..,K}{\mathrm{argmax}}\{\mathrm{p}(k|D)\},\end{aligned} \tag{9}$$

where $\mathrm{p}(k|D)$ denotes the posterior probability of topic k returned by LDA given as input the sentence/document D and $x'_k(w_d)$ is the mapped representation of word w_d for topic k. For the case of paraphrase identification, we extract a single feature vector for each sentence-pair via concatenating the features of the individual sentences.

After feature extraction, we train a linear Support Vector Classifier (SVM) (Pedregosa et al., 2011) using the proposed train/test sets for both tasks. We report the best results for each experimental configuration after tuning the SVM's penalty parameter of the error term using 500-dimensional word embeddings.

5 Results

In Table 1 we compare our model (UTDSM) with our baseline (Global-DSM) and other state-of-the-art multi-prototype approaches for the contextual semantic similarity task. It is clear that all different setups of UTDSM perform better than the baseline for both contextual semantic similarity metrics. Using a single Gaussian distribution (UTDSM + GMM (1)) at the smoothing step of our method produces similar results to the baseline model. This is anticipated as both methods provide a centroid representation of a word's diverse semantics. In terms of $\mathrm{MaxSimC}$ the model consistently yields higher performance when the list of semantic anchors is induced via our unsupervised method instead of using randomly selected anchor words (UTDSM Random). We also observe that random anchoring performs slightly worse than UTDSM with respect to $\mathrm{AvgSimC}$. This result validates our hypothesis that the representations of words, which share consistent similarity distributions across different topic domains, constitute informative *semantic anchors* that determine the mappings between semantic vector spaces.

Method	AvgSimC	MaxSimC
Liu et al. (2015a)	67.3	68.1
Liu et al. (2015b)	69.5	67.9
Amiri et al. (2016)	70.9	-
Lee and Chen (2017)	68.7	67.9
Guo et al. (2018)	69.3	68.2
300-dimensions		
Global-DSM	67.1	67.1
UTDSM Random	69.1 ± 0.1	66.4 ± 0.2
UTDSM	**69.6**	67.1
UTDSM + GMM (1)	67.4	67.4
UTDSM + GMM (2)	68.4	**68.3**
UTDSM + GMM (3)	68.9	**68.3**
UTDSM + GMM (8)	69.1	68.0
UTDSM + GMM (10)	69.0	67.8
500-dimensions		
Global-DSM	67.6	67.6
UTDSM Random	69.4 ± 0.1	66.5 ± 0.3
UTDSM	**70.2**	68.0
UTDSM + GMM (1)	67.6	67.6
UTDSM + GMM (2)	68.8	**68.6**
UTDSM + GMM (3)	69.0	68.5
UTDSM + GMM (8)	69.5	68.5
UTDSM + GMM (10)	69.2	68.0

Table 1: Performance comparison between different state-of-the-art approaches on SCWS, in terms of Spearman's correlation. UTDSM refers to the projected cross-topic representation, UTDSM Random refers to the case when random words served as anchors and GMM (c) corresponds to GMM smoothing with c components.

Furthermore, we observe that GMM smoothing has a different effect on the $\mathrm{MaxSimC}$ and $\mathrm{AvgSimC}$ metrics. Specifically, for $\mathrm{AvgSimC}$ we consistently report lower results when GMM smoothing is applied for different number of components. We attribute this behavior to a possible loss of model capacity—decrease in the number of topic embeddings—that is capable of capturing additional topic information. At the same time, our smoothing technique highly improves the performance of $\mathrm{MaxSimC}$ for all possible configurations. Given that this metric is more sensitive to noisy word representations, this result indicates that our technique lessens the noise introduced to our system and captures finer-grained topic senses of words.

Overall, the performance of our model is highly competitive to the state-of-the-art models in terms of $\mathrm{AvgSimC}$, for 500-dimensional topic embeddings. We also achieve state-of-the-art performance for the $\mathrm{MaxSimC}$ metric, using smoothed topic embeddings of 300 or 500 dimensions with 2 or 3 Gaussian components.

Method	Precision	Recall	F1-score	Accuracy
LDA	39.7	41.8	38.8	41.8
Global-DSM	62.9	63.3	62.9	63.3
$\mathrm{MaxC_D}$	61.9	63.0	62.0	63.0
$\mathrm{Avg_D}$	63.5	64.6	63.3	64.3
$\mathrm{AvgC_D}$	**64.6**	**65.5**	**64.5**	**65.5**

Table 2: Evaluation results of multi-class text classification.

Evaluation results on text classification are presented in Table 2. We observe that our model performs better than the baseline across all metrics for both averaging approaches ($\mathrm{AvgC_D}$, $\mathrm{Avg_D}$), while the usage of dominant topics appears to have lower performance ($\mathrm{MaxC_D}$). Specifically, we get an improvement of $2 - 2.5\%$ on topic-based average and $0.5 - 1\%$ on simple average combination compared to using Global-DSM.

Method	Precision	Recall	F1-score	Accuracy
Global-DSM	68.6	69.2	62.0	69.2
$\mathrm{MaxC_D}$	**69.0**	69.3	62.1	69.3
$\mathrm{Avg_D}$	67.7	**69.4**	**64.0**	**69.4**
$\mathrm{AvgC_D}$	68.8	69.4	62.6	69.4

Table 3: Evaluation results on paraphrase detection task.

Results for the paraphrase identification task are presented in Table 3. $\mathrm{Avg_D}$ yields the best results especially in F1 metric showing that cross-topic representations are semantically richer than single embeddings baseline (Global-DSM). Although we apply the topic distributions $\mathrm{p}(k|D)$ extracted from LDA (document-level model) to a sentence-level task, improvements over the baseline are also shown in the $\mathrm{AvgC_D}$ and $\mathrm{MaxC_D}$ cases.

Overall, the proposed UTDSM model outperforms the baseline Global-DSM model on both semantic similarity and downstream tasks.[7]

6 Cross-Domain Semantic Analysis

Finally, we carry out a cross-domain semantic analysis to detect the variations of a word's meaning in different topic domains. To that end, we use a list of known polysemous words and measure the semantic similarity between different topic representations of the same ambiguous word. The ultimate goal of this analysis is to validate that our model captures known thematic variations in semantics of polysemous words.

Table 4 includes examples of our analysis. The most probable words of the topics (second column) give an intuitive sense of their major contexts, while their nearest neighbors (third column) infer the sense of the target word in the corresponding topic domain. For example, the word *drug* is mostly related to "medication" in a broad medical domain; it experiences though a slight shift from this meaning when it resides in a topic about "illegal substances". Furthermore, the highly polysemous word *act* shifts from meaning "statute" to meaning "performance" under the corresponding law and art topics. Similar semantic variations are observed for words *python*, *rock* and *nursery*.

Moreover, in Figure 3 we visualize the topic embeddings of seven words before and after projecting the topic-based DSMs to the unified space, using principal component analysis. We additionally depict the Gaussian distribution learned from the topic representations of each word reflecting the uncertainty of their meanings. The center of each distribution is specified by the mean vector and contour surface by the covariance matrix. On the left, we depict the position of words prior to applying the unsupervised mapping approach where the topic sub-spaces are unaligned. In the unaligned space, words demonstrate similar area coverage regardless of their polysemy. After the mappings, we see on the right that the area under a word's distribution is indicative of its degree of polysemy. Specifically, we observe that the variance of the learned representations becomes larger for the cases of polysemous words such as

[7] Similar results were obtained for each metric using smoothed word embeddings. Also, there are no standard evaluation approaches for comparison of previous works on downstream tasks.

[8] Note that a topic domain is described as a distribution over words in our model.

Word	Topic Words	Nearest Neighbors	Similarity
drug	health, medical, cancer, treatment, disease	insulin, therapy, heparin, chemotherapy, vaccines	0.61
	drug, health, marijuana, alcohol, effects	meth, cocaine, methamphetamine, mdma, heroin	
act	law, court, legal, tax, state	bylaw, legislature, complying, entities, entitlement	0.39
	music, guitar, piano, dance, theatre	touring, pantomime, weekend, shakespeare, musical	
python	garden, plant, fish, bird, animal	macaw, crocodile, hamster, albino, rattlesnake	0.27
	software, forum, download, windows, web	algorithm, parser, notepad, gui, tutorial	
rock	mountain, river, park, road, trail	geology, slab, limestone, waterfalls, canyon	0.43
	music, guitar, piano, dance, theatre	touring, acoustic, americana, songwriter, combo	
nursery	garden, plant, tree, flower, gardening	camellias, succulents, greenhouse, ornamental, grower	0.46
	university, school, college, education, program	prep, montessori, grammar, preschool, infant	

Table 4: Examples of polysemous words and the change of meaning between different topic domains. First column lists the example target words. Second column includes the most probable words of the topic domains[8] these words are assigned to. Each row corresponds to a different topic domain. Third column shows the nearest monosemous neighbors of the target word in the corresponding topic domain. The last column corresponds to the cosine similarity between the two topic representations of the target word.

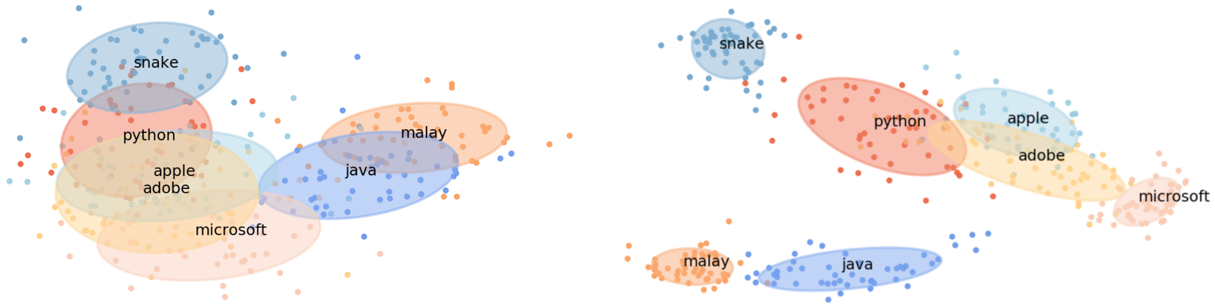


Figure 3: A 2-dimensional projection of the latent semantic space encoded in our unified vector space model, depicting the topic word representations of 7 words before (left) and after (right) mapping the TDSMs to the global semantic space.

"python", "java", "adobe" in order to assign some probability to their diverse meanings. Monosemous words such as "snake", "microsoft" and "malay" have smaller variances. Furthermore, we observe that the semantic relationships between words are much better captured by their corresponding positions in the aligned space.

7 Conclusion

We present an unsupervised approach of mapping multiple topic-based DSMs to a unified vector space in order to capture different contextual semantics of words. We assume that words having consistent similarity distributions regardless of the domain they exist in could be considered informative semantic anchors that determine the mappings between semantic spaces. The projected word embeddings yield state-of-the-art results on contextual similarity compared to previously proposed unsupervised approaches for multiple word embeddings creation, while they also outperform single vector representations in downstream NLP tasks. In addition, we provide insightful visualizations and examples that demonstrate the capability of our model to capture variations in topic semantics of words.

As future work, one can hypothesize that the area a word covers in the mapped space reveals its semantic range. In this direction, a refinement of the semantic anchor selection approach could be explored in an iterative way assuming that the variance of a word's Gaussian distribution denotes its degree of polysemy (Vilnis and McCallum, 2015). Moreover, we would like to explore a more sophisticated smoothing technique where the number of Gaussian components is adapted for each word. Given that Gaussian mixture embeddings could capture the uncertainty of a word's representation in the semantic space one could also investigate different metrics for measuring the semantic relationship between word pairs that go beyond their point-wise comparison. Finally, it may be helpful to investigate non-linear mappings between semantic spaces using deep neural network architectures.

Acknowledgments

Thi work has been partially funded by the BabyRobot project, supported by the EU Horizon 2020 Program undergrant #687831.

References

Hadi Amiri, Philip Resnik, Jordan Boyd-Graber, and Hal Daumé III. 2016. Learning text pair similarity with context-sensitive autoencoders. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 1882–1892.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proc. Conference on Empirical Methods in Natural Language Processing*, volume 1, pages 2289–2294.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035.

Fenia Christopoulou, Eleftheria Briakou, Elias Iosif, and Alexandros Potamianos. 2018. Mixture of topic-based distributional semantic and affective models. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 203–210.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*.

Fenfei Guo, Mohit Iyyer, and Jordan Boyd-Graber. 2018. Inducing and embedding senses with scaled gumbel softmax. *arXiv preprint arXiv:1804.08077*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1489–1501.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics*.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2–3):146–162.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 873–882.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 95–105.

Guang-He Lee and Yun-Nung Chen. 2017. Muse: Modularizing unsupervised sense embeddings. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pages 327–337.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1722–1732.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2015a. Learning context-sensitive word embeddings with neural tensor skip-gram model. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI*, pages 1284–1290.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015b. Topical word embeddings. In *Proc. AAAI Conference on Artificial Intelligence*, pages 2418–2424.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. volume abs/1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *HLT-NAACL*.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient nonparametric estimation of multiple embeddings per word in vector space. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069.

Dai Quoc Nguyen, Dat Quoc Nguyen, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2017. A mixture model for learning multi-sense word embeddings.

In *Proc. 6th Joint Conference on Lexical and Computational Semantics*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Mohammad Taher Pilehvar and Nigel Collier. 2016. De-conflated semantic representations. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1680–1690.

Victor Prokhorov, Mohammad Taher Pilehvar, Dimitri Kartsaklis, and Nigel Collier. 2017. Learning rare word representations using semantic bridging.

Joseph Reisinger and Raymond Mooney. 2010. Mixture Model with Sharing for Lexical Semantics. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1182.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1793–1803.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, pages 627–633.

Peter H. Schönemann. 1966. *A generalized solution of the orthogonal procrustes problem*.

Amr Sharaf, Shi Feng, Khanh Nguyen, Kiante Brantley, and Hal Daumé III. 2017. The umd neural machine translation systems at wmt17 bandit learning task. In *Proceedings of the Second Conference on Machine Translation*, pages 667–673.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *Proc. 53rd Annual Meeting of the Association for Computational Linguistics*, 1:1556–1566.

Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A probabilistic model for learning multi-prototype word embeddings. In *Proc. International Conference on Computational Linguistics (COLING)*, pages 151–160.

Peter D Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44:533–585.

Luke Vilnis and Andrew McCallum. 2015. Word representations via gaussian embedding. In *International Conference on Learning Representations (ICLR)*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.

Xiaoqing Zheng, Jiangtao Feng, Yi Chen, Haoyuan Peng, and Wenqing Zhang. 2017. Learning context-specific word/character embeddings. In *Proc. AAAI Conference on Artificial Intelligence*, pages 3393–3399.