Mixture of Topic-based Distributional Semantic and Affective Models

Fenia Christopoulou, Eleftheria Briakou, Elias Iosif, Alexandros Potamianos
School of ECE, National Technical University of Athens, Zografou 15780, Athens, Greece
fenia.christopoulou@gmail.com, el12145@central.ntua.gr, iosife@central.ntua.gr, potam@central.ntua.gr

***Abstract*—Typically, Distributional Semantic Models (DSMs) estimate semantic similarity between words using a single-model, where the multiple senses of polysemous words are conflated in a single representation. Similarly, in textual affective analysis tasks, ambiguous words are usually not treated differently when estimating word affective scores. In this work, a semantic mixture model is proposed enabling the combination of word similarity scores estimated across multiple topic-specific DSMs (TDSMs). Based on the assumption that semantic similarity implies affective similarity, we extend this model to perform sentence-level affect estimation. The proposed model outperforms the baseline approach achieving state-of-the-art results for semantic similarity estimation and sentence-level polarity detection.**

I. Introduction

Distributional semantic models (DSMs) aim at representing the meaning of lexical entities by encoding linguistic features extracted from text corpora. Word-level representations are the building block for more complex phrase– and sentence–level representations used for similarity computation [1], [2].

Word-level DSMs can be broadly categorized, with respect to the extraction of contextual features, into unstructured and structured. The bag-of-words model is the most widely used approach, lacking however some desirable characteristics such as "order sensitivity" [3]. Unlike unstructured models, the order of extracted features is taken into account in the framework of structured DSMs via the exploitation of syntactic relationships (e.g., argument structures and modifications) [4]. Recently, the computation of contextual features was posed in a learning-based framework, where the goal is to estimate the context in which the words of interest are expected to occur [5], [6].

The multiple senses of polysemous words are typically not directly encoded in DSMs. To address this issue, exemplar models were proposed, where the meaning of a word was represented by a set of stereotypical corpus sentences instead of a single feature vector [7]. An alternative approach is the use of topic modeling, which results into a parsimonious representation of the topics (thematic domains) that exist in the corpus under analysis. Typically, each topic is represented as a distribution of words being salient for the respective topic. Latent Dirichlet Allocation (LDA) [8] constitutes the most widely-used topic modeling approach using models proposed in [9] and [10]. Extensions of LDA include the Correlated Topic Model (CTM) [11] and the Pachinko Allocation machine [12] which aim to improve the topic detection process by measuring the correlation between topics. The main motivation behind the use of topic models, for the task of word semantic similarity computation, is to adapt the similarity estimates provided from various topics. This is similar to using semantic mixture models to encode multiple senses in words.

In this work, a topic-based semantic mixture model is proposed for the computation of semantic similarity between words. This is motivated by previous approaches (e.g., [13]–[15]) to utilize a combination of similarities computed via topic-based DSMs. In addition, the proposed mixture model is incorporated into a semantic-affective mapping used for estimating affective scores for sentences. Significant improvements over the baseline models were achieved reaching state-of-the-art performance.

II. Related Work

Sense-agnostic representations were introduced in [16] where context-dependent clusters were combined to semantically represent words. The model was extended in [17] to automatically estimate the number of clusters.

Subsequent approaches mostly relied on neural network architectures that encode multi-sense information. The work of [18] proposed a skip-gram word2vec model that combined global and local context information to train word embeddings. They used spherical k-means to cluster word context, assuming a fixed number of possible senses per word. Later, the skip-gram sense-embeddings were refined through backpropagation, as described in [19]. An improved version of the same network was introduced by [20], where the posterior probability of a context word was represented as a mixture of the senses of the target word. In [21] the skip-gram model was further modified to jointly train word and sense vectors, while WordNet glosses were used to assign senses to the target words. A different approach proposed by [22] utilized bilingual resources to learn multiple sense-specific embeddings for each ambiguous word, with a recurrent neural network. The model described in [23] considered a Gaussian mixture for each word, where each Gaussian component represented a word sense. A dynamic skip-gram mixture model was proposed, able to detect different number of senses for each word during training. Additionally, skip-gram was extended in [14] to simultaneously train word and topic embeddings and to identify their interactions. The representation of each word was formed as a mixture of the word's different senses and a topic embedding was produced by averaging the word embeddings under the topic. Finally, LDA was employed into

the skip-gram model, as depicted in [13], to get the distribution of a word over the topics.

Different techniques involved knowledge-based approaches that use sense inventories to obtain word-sense embeddings [24], ontologically grounded senses [25], WordNet lexemes, where each word is considered the sum of its lexemes [26] and Wikipedia links to identify specific senses [27]. A model for training multiple embeddings per word according to its senses, based on the Chinese restaurant process is described in [28]. The proposed approach followed the idea that a word should have a new sense if there is corresponding evidence in the context. State-of-the-art models, that deal with contextual word similarity, make use of context auto-encoders for each word [29] or neural networks for joint extraction of words and contexts [30]. The best performance was achieved in [31], where pre-trained word representations were linked to WordNet. Biased words were utilized towards the target word to find the minimum distance among them and considered this embedding as a sense-agnostic embedding.

Topic models were also recently used for sentiment analysis tasks. In [32] both topics and sentiments were detected in Weblogs using a Topic-Sentiment mixture model. Multinomial distributions were incorporated based on the assumption that a document contains different topics and each topic consists of different sentiments. Similarly, in [33] a joint model of sentiments and topics was proposed. The LDA algorithm was modified in order to consider sentiment labels for a document. Another model introduced in [34] analyzed how sentiments are expressed for different aspects, by assuming that all words in a single sentence are generated from one aspect. Two sentiment topic models were proposed in [35] to associate latent topics with evoked emotions of readers. Finally, a topic-based affective mixture model [15] predicted the polarity of tweets by training topic-specific Support Vector Machines.

III. Semantic Mixture Models

The typical use of DSMs deals with the creation of a single feature space, where the multiple senses of a polysemous word (assuming a generic corpus of wide coverage) are conflated into a single semantic representation (*sense-agnostic DSMs*). In this framework, the computation of semantic similarity between a pair of words is performed across all of their senses that appear in the corpus. For various semantic tasks related to similarity computation such models were found to achieve very good performance despite their divergence from the *maximum sense similarity* assumption. This assumption suggests that the semantic similarity between two words can be estimated as the similarity of their two closest senses [36].

In this work, the aforementioned assumption is adopted via the creation of topic-based sub-corpora with respect to any pair of words, w_i and w_j, subjected to similarity computation. The goal is the words of the pair to co-occur in each sub-corpus with their closest senses, pertaining to the relevance with the respective topics. This approach is different compared to the typical corpus-based word sense induction (also referred to as sense discovery) [37], where the discovery is performed individually for each word. The similarity between w_i and w_j is computed by a mixture model that combines similarity scores computed over multiple topic-based sub-corpora. The steps of the proposed approach are briefly described next.

A. Topic Modeling

The Latent Dirichlet Allocation (LDA) algorithm [8] is a generative process that attempts to identify possible topics (thematic domains) residing in a corpus. The underlying assumption of the algorithm is that a document collection can be represented as a probabilistic mixture of a fixed number of topics, where each topic is a distribution over the words in the collection. A trained topic model produces a distribution of words for each topic, that are semantically related under the corresponding topic. In the proposed approach, the possible topics that occur in a corpus are identified by training the LDA algorithm on the underlying corpus, for a number of topics T.

B. Creation of Topic-based Sub-corpora

In order to topically adapt the semantic space by training topic DSMs (TDSMs), in-domain sub-corpora need to be created. The isolation of different word senses is achieved by collecting topic-related snippets into separate bodies of text, using the trained topic model.

In more detail, the model is applied on the sentences of a corpus. This choice adheres to the basic principles of topic modeling, since sentences are topically complete and coherent units. As a result, each sentence is probabilistically associated with a list of topics, discussed in the sentence, according to the topic model. A sub-corpus is created for each topic $t \in T$ by aggregating the sentences, the posterior probabilities of which are maximized for t. This hard-clustering scheme may result in sub-corpora of limited size. In order to relax this limitation, a soft-clustering scheme is adopted. Specifically, a sentence is allowed to be included in a topic-specific corpus when the posterior probability for the corresponding topic exceeds a threshold h. Sentences exhibiting equal posterior probabilities across all topics are excluded from this process, as considered too generic to provide any topic-related information.

C. Semantic Similarity Computation

The topic-based semantic representations of words are produced by training a DSM on each sub-corpus that resulted from the previous step. We define L_T as the set of T topic DSMs (TDSMs) derived from the LDA algorithm, where λ_t is the DSM trained on topic t out of the T topics in total.

The semantic similarity between two words w_i and w_j is computed using different similarity metrics with respect to the presence of context for each pair. A mixture model of topic-based semantic similarities is incorporated to produce the final similarity $S(w_i, w_j)$ between a word pair. In accordance with [17], we define two non-contextual metrics:

$$S_{\text{AvgSim}}(w_i, w_j; L_T) = \frac{1}{T} \sum_{t=1}^{|T|} S_t(w_i, w_j; \lambda_t), \qquad (1)$$

$$S_{\text{MaxSim}}(w_i, w_j; L_T) = \max_{t \in T}\{S_t(w_i, w_j; \lambda_t)\}, \quad (2)$$

where $S_t(w_i, w_j; \lambda_t)$ is the semantic similarity of w_i and w_j computed by the λ_t DSM, which was built using the sub-corpus that corresponds to topic t. In AvgSim (1), the unweighted average of all topic-based pairwise semantic similarities is computed. In (2) only the maximum pairwise similarity, among T topics, is selected.

When context information is provided for a pair, a shared context $c = c(w_i) \oplus c(w_j)$ is formulated by concatenating the contexts of each word $c(w_i)$ and $c(w_j)$. The topic model is fed with c and outputs a list of candidate topics for c, along with the corresponding posterior probabilities $p(t|c)$. These topics are utilized for identifying the respective sub-corpora, which are used to train topic-specific DSMs (TDSMs).

In order to consider context information, AvgSim (1) and MaxSim (2) are modified. Similarly to [17], we define two more detailed similarity metrics[1]:

$$S_{\text{AvgSimC}}(w_i, w_j; L_T) = \frac{\sum_{t=1}^{|K(c)|} p(t|c)\ S_t(w_i, w_j; \lambda_t)}{\sum_{t=1}^{|K(c)|} p(t|c)}, \quad (3)$$

$$S_{\text{MaxSimC}}(w_i, w_j; L_T) = S_{\hat{t}}(w_i, w_j; \lambda_{\hat{t}})$$
$$\hat{t} = \arg\max_{t \in K(c)}\{p(t|c)\}, \quad (4)$$

where $K(c)$ are the candidate topics returned by the topic model with a posterior probability larger than 0.01, when given as input a shared context c, $p(t|c)$ denotes the posterior probability of topic t for c, while $S_t(w_i, w_j; \lambda_t)$ is the semantic similarity of w_i and w_j from the DSM that corresponds to topic t. Because the number of candidate topics can be less or equal to the total number of topics ($K(c) \leq T$), for which LDA is trained, the posterior probabilities are normalized to sum to unity.

Given c as input to the topic model, (3) computes a weighted average of topic-based semantic similarities using the topics posterior probabilities as weights[2]. The model takes the middle road between the maximum sense similarity hypothesis and the sense-agnostic DSMs. This hypothesis is adopted for the identification of sub-corpora in which w_i and w_j appear with related senses under the thematic domain of the corresponding topic. The incorporation of the mixture weights in the computation of the final similarity relaxes the hypothesis. Using (4) a pair is assigned the semantic similarity of the topic with the maximum posterior probability, hence the dominant topic in the provided context.

Additionally, we introduce a fusion model that combines information from multiple topic models trained for different number of topics. In more detail, for a topic model trained on T topics, the semantic similarity of a word pair is calculated using one of the aforementioned metrics, as defined in (1)−(4). Among the similarities produced by training the topic model for various number of topics, we select the maximum pair similarity over a group G, of topic DSM sets L_T, generated by different topic models:

$$S_{\text{Fuse}}(w_i, w_j) = \max_{L_T \in G}\{S_{*\text{Sim}}(w_i, w_j; L_T)\}, \quad (5)$$

where $S_{*\text{Sim}}(w_i, w_j; L_T)$ is the w_i, w_j pair similarity computed with (1)−(4), using a topic model trained on T topics and G is the group of DSM sets that will be fused.

Finally, we employ a linear regression model to combine pairwise similarities between topic DSMs (TDSMs), resulted from a topic model trained on T topics. The model aims to minimize the Mean Squared Error (MSE) by training a set of β weights on a group of similarities between words. The motivation behind this idea is to learn how to combine topic-specific similarities for isolated words. The context-dependent similarity metric (3) requires additional input (context) to estimate how much each topic-similarity will be weighted. In contrast, when no context is present, instead of assuming that all topics contribute equally to the estimation of a pairwise similarity, as described in (1), we argue that a linear combination of topic-similarities will produce a more precise estimation,

$$S_{\text{LRSim}}(w_i, w_j; L_T) = \beta_0 + \sum_{t=1}^{|T|} \beta_t\ S_t(w_i, w_j; \lambda_t), \quad (6)$$

where β_t are learned weights by the regression model for the corresponding topic t, $S_t(w_i, w_j; \lambda_t)$ is the similarity of a pair w_i, w_j computed from the DSM trained on the sub-corpus of topic t and β_0 is a bias weight. The β weights sum to unity.

IV. Affective Analysis of Text

We extend the semantic mixture model to predict affective scores for sentences. For this purpose we use the semantic-affective model (SAM) proposed in [38]. The model exploits the continuous affective space (valence-arousal-dominance) and computes affective ratings for unknown words, as shown in Fig. 1. The required inputs to the model are i) a set of words with known affective scores, named an affective lexicon, ii) a DSM trained on a general-purpose corpus and iii) a mapping from the semantic to the affective space in the form of trainable weights.

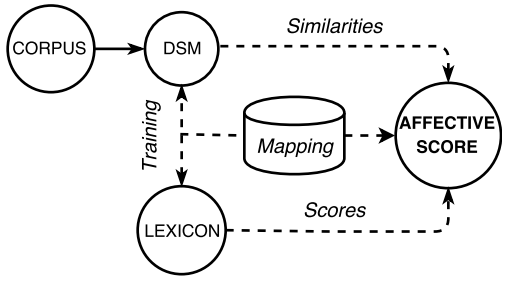

Fig. 1. Representation of the semantic-affective model.

[1]The additional capital letter C stands for *Context*.

[2]For pairs that share the same word, but are found in different contexts, the model always assigns them a similarity score equal to one, as their representations are extracted from the same topic-based DSM.

A. Semantic-Affective Model

1) Word-level scores: The words affective scores are estimated using the semantic-affective model directly. Firstly, the model selects a subset of representative words from the affective lexicon that correspond to words with extreme affective ratings, named seed words. Then, the affective score for an unknown word is generated as a linear combination of the seed words affective scores and their similarity to the unknown word, weighted by trainable weights,

$$v(w_j) = \alpha_0 + \sum_{n=1}^{N} \alpha_i \; v(s_i) \; S(s_i, w_j; \lambda), \tag{7}$$

where w_j is the word whose affective score we aim to estimate, $s_1, ..., s_N$ are the seed words, α_i is the weight corresponding to seed word s_i, $v(s_i)$ is the valence rating for seed word s_i and $S(s_i, w_j; \lambda)$ is the semantic similarity between seed word s_i and unknown word w_j from a DSM λ. The α weights are learned by Mean Squared Error minimization via Ridge Regression on (7).

2) Sentence-level scores: Based on the principle of compositionality [39] we can estimate the meaning of a sentence as the sum of the meaning of its parts. Consequently, the affective score of a sentence can be computed using three word-based fusion models, as proposed in [38]:

i) Linear Fusion

$$v(s) = \frac{1}{N} \sum_{i=1}^{N} v(w_i) \tag{8}$$

ii) Weighted Fusion

$$v(s) = \frac{1}{\sum_{i=1}^{N} |v(w_i)|} \sum_{i=1}^{N} v(w_i)^2 \; \mathrm{sgn}(v(w_i)) \tag{9}$$

iii) Non-linear Max Fusion

$$\begin{aligned} v(s) &= \max_i \{|v(w_i)|\} \; \mathrm{sgn}(v(w_z)) \\ w_z &= \arg\max_i \{|v(w_i)|\}, \end{aligned} \tag{10}$$

where $v(w_i)$ is the valence score of word w_i, N is the total number of words in a sentence and $\mathrm{sgn}(x)$ is the signum function. Linear fusion, defined in (8), equally weights the affective scores of the words in a sentence to produce a sentence score. Weighted fusion, defined in (9), weights more words with higher absolute affective scores and max fusion, defined in (10), considers only the word w_z with the maximum absolute affective score in the sentence.

3) Affective Mixture Model: Each sentence s, the affective score of which we aim to estimate, is given as input to a trained topic model. A list of candidate topics and posterior probabilities is produced, based on the likelihood of the topics being discussed in the sentence. A mixture model is used to estimate the similarities between the words of the sentence (unknown words w_j) and a set of seed words (known words s_i). The topic similarities are incorporated into (7) to estimate sentence words affective scores, as shown in (11). Finally, using a word fusion scheme (8)−(10) we compute the affective rating of the sentence,

$$v_{\text{adapt}}(w_j) = \alpha_0 + \sum_{n=1}^{N} \alpha_i \; v(s_i) \; S_{\text{AvgSimC}}(s_i, w_j; L_T), \tag{11}$$

where $v(s_i)$ is the valence score of seed word s_i, α_i is the weight of seed word s_i, $S_{\text{AvgSimC}}(s_i, w_j; L_T)$ is the adapted semantic similarity between seed word s_i and sentence word w_j, as resulted from (3), and $v_{\text{adapt}}(w_j)$ is the final adapted valence score for a sentence word. The use of AvgSimC is based on its good performance reported in the literature.

V. Experiments and Evaluation

In this section, the experimental settings along with the evaluation results are briefly presented.

A. Experimental Settings

1) Corpora: We used two generic corpora in English to train the LDA-based model and create the topic-specific sub-corpora. First, we used a web-harvested corpus (Web) consisting of 116 million sentences created as follows [40]: Starting from a lexicon, an individual query was formulated for each lexicon entry and the 1000 top ranked results (document snippets) were retrieved and aggregated using the Yahoo! search engine. Second, we used the English Wikipedia[3], containing 8.5 million articles. During the training of the topic model, we used the articles found in the Wikipedia corpus, while for the Web corpus we used pseudo-documents constructed as groups of snippets retrieved by the same search query (i.e., the grouped snippets are topically related).

2) Topic-modeling: The Gensim Toolbox [41] was used for topic modeling based on LDA. We experimented with up to 100 topics, with 200 model iterations, while the rest of Gensim parameters were fixed to their default values.

3) Topic-specific sub-corpora: Regarding the creation of topic-specific sub-corpora (described in Section III-B), the individual sentences were classified for each corpus adopting the soft-clustering scheme[4]. Approximately 90 million sentences were extracted from the Wikipedia articles using the University of Illinois sentence segmentation tool[5]. The h threshold used in this process was set to 0.1 after an empirical analysis of the created sub-corpora.

4) DSMs: All the DSMs in this work were created using Google's implementation of word2vec[6] and the Continuous Bag-of-Words (CBOW) approach for the extraction of contextual features. We built two baseline DSMs using the aforementioned corpora without applying topic modeling. We used the baseline models for setting the number of dimensions of the feature space (300 and 500 for Web and Wikipedia corpora, respectively) and the size of the context window (five for both corpora). This was done with respect to various datasets

[3]https://dumps.wikimedia.org/enwiki/20160720/

[4]The hard-clustering approach was found to yield lower performance compared to the soft-clustering one.

[5]https://cogcomp.cs.illinois.edu/page/tools_view/2

[6]https://code.google.com/archive/p/word2vec/

dealing with word semantic similarity. The default settings were used for all other word2vec parameters. These parameters were fixed for each corpus and used during the training of the respective topic-specific DSMs (TDSMs).

5) Semantic similarity: The word similarities incorporated in (1)−(6) were computed by taking the cosine of their respective vectors, which correspond to the word embeddings computed by word2vec for each sub-corpus. For the Linear Regression model (6), we used the Leave-One-Out method on MEN dataset with 2000 pairs for training and 1000 pairs for testing. In order to test on the WS-353 dataset, the entire MEN dataset was used for training. The performance of the proposed topic-based mixture model was evaluated for the task of word similarity computation with and without context information, on the datasets described in Table I.

TABLE I
DATASETS USED FOR IN-CONTEXT AND OUT-OF-CONTEXT SEMANTIC SIMILARITY COMPUTATION.

Dataset	Pairs	Similarity Range	Context
MEN [42]	3000	$[0, 50]$	no
WS-353 [43]	353	$[0, 10]$	no
SCWS [18]	2003	$[0, 10]$	yes

For the datasets that provide words in isolation (MEN and WS-353) the MaxSim metric (2) is reported. For SCWS dataset, context-dependent metrics are reported (3) and (4), along with the AvgSim out-of-context metric (1) in order to compare with the literature. To experiment with the fusion model (5) we used different combinations of topic groups, from 5 to 100 topics. The Spearman correlation between the automatically computed similarity scores and the human similarity ratings is the evaluation metric for all datasets.

6) Affective Model: The semantic-affective model requires an affective lexicon. We selected the Affective Norms for English Words (ANEW) [44] similarly to [45]. The mapping from the semantic to the affective space was computed with 600 seed words using similarities from a global DSM trained with word2vec on the entire Web corpus. We grounded all negative semantic similarities to zero and applied MSE Ridge Regression to train the α weights.

In order to evaluate the proposed model we used the SemEval 2007 Task 14 dataset [46]. The dataset includes 250 annotated sentences for training and 1000 sentences for testing, from news headlines. Each sentence is associated with a sentiment score, in the valence dimension, in range $[-100, 100]$ which was rescaled to $[-1, 1]$ and represents scores from highly negative to highly positive headlines. We measured the Spearman correlation ρ score with ground truth valence scores provided by the dataset. We used only the Web-based corpus for these experiments considering its performance for the out-of-domain datasets in the semantic similarity task.

B. Evaluation Results

1) Semantic Similarity: In Table II the performance of the proposed approach, topic-specific DSMs (TDSMs) (1)−(4) and the corresponding variations, TDSMs-Fuse (5) and TDSMs-LR (6), is presented for different datasets and corpora. Additionally, the performance of a baseline model that utilizes a single topic (No Topics) is reported. The models are compared with various approaches proposed in the literature.

TABLE II
PERFORMANCE COMPARISON BETWEEN DIFFERENT APPROACHES AND DATASETS FOR SEMANTIC SIMILARITY COMPUTATION, IN TERMS OF SPEARMAN'S ρ CORRELATION.

Approach	*Out-of-Context*		*In-Context*		
	WS-353	MEN	SCWS		
			MaxSimC	AvgSim	AvgSimC
[18]	0.713	–	–	0.628	0.657
[20]	–	–	0.636	–	0.654
[21]	–	–	–	0.662	0.689
[19]	0.709	–	–	0.673	0.693
[23]	0.678	–	0.536	0.646	–
[24]	**0.779**	0.805	0.589	–	0.624
[25]	0.639	0.646	–	–	0.657
[28]	–	–	–	–	0.697
[27]	0.739	–	0.662	–	0.664
[14]	–	–	0.679	–	0.695
[13]	–	–	0.673	–	0.681
[26]	–	–	–	0.689	0.698
[31]	–	0.786	–	**0.708**	**0.715**
[29]	–	–	–	–	0.709
[30]	–	–	–	–	0.699
Web Corpus					
TDSMs	0.722	0.800	0.678	0.678	0.702
TDSMs-Fuse	–	–	0.674	0.6764	0.705
TDSMs-LR	0.727	**0.838**	–	–	–
No Topics	0.703	0.773	0.659		
Wikipedia Corpus					
TDSMs	0.698	0.753	**0.683**	0.696	0.701
TDSMs-Fuse	–	–	0.6814	0.685	0.707
TDSMs-LR	0.695	0.796	–	–	–
No Topics	0.644	0.731	0.669		

The proposed topic-based models (TDSMs) outperform the baseline for all datasets and corpora. For the MaxSimC metric (2) the model achieves the best performance (0.683), regarding the SCWS dataset. For the other two metrics, AvgSim (1) and AvgSimC (3), the proposed approach achieves 0.696 and 0.702 correlation being close to the top performing systems (0.708 and 0.715, respectively). The fusion model (5) further improves the performance of the TDSMs model for the AvgSimC metric in both corpora (0.705 and 0.707, respectively).

Concerning the out-of-context datasets, the Linear Regression model (TDSMs-LR) achieves state-of-the-art performance (0.838 correlation) for MEN dataset exceeding all models proposed in the literature. The same approach ranks third (0.727) compared to the top performing model of [24], regarding WS-353.

Fig. 2a and 2b illustrate the performance of the TDSMs-LR model, as a function of the number of topics. Regarding the

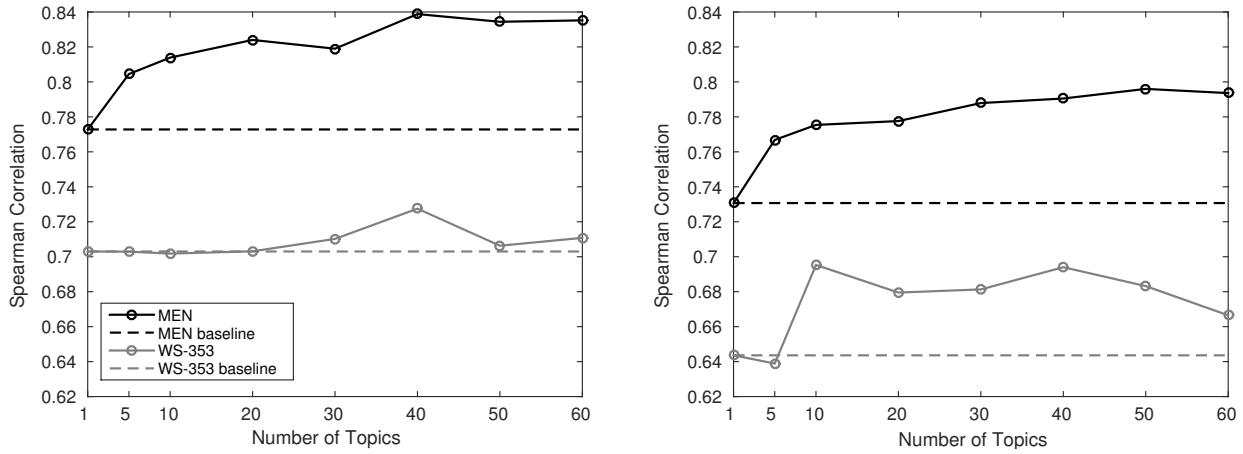

Fig. 2. Spearman ρ correlation for MEN and WS-353 datasets as a function of the number of topics using the Linear Regression Topic DSMs model (TDSMs-LR): (a) Web corpus, (b) Wikipedia corpus. The baseline corresponds to a model with a single topic.

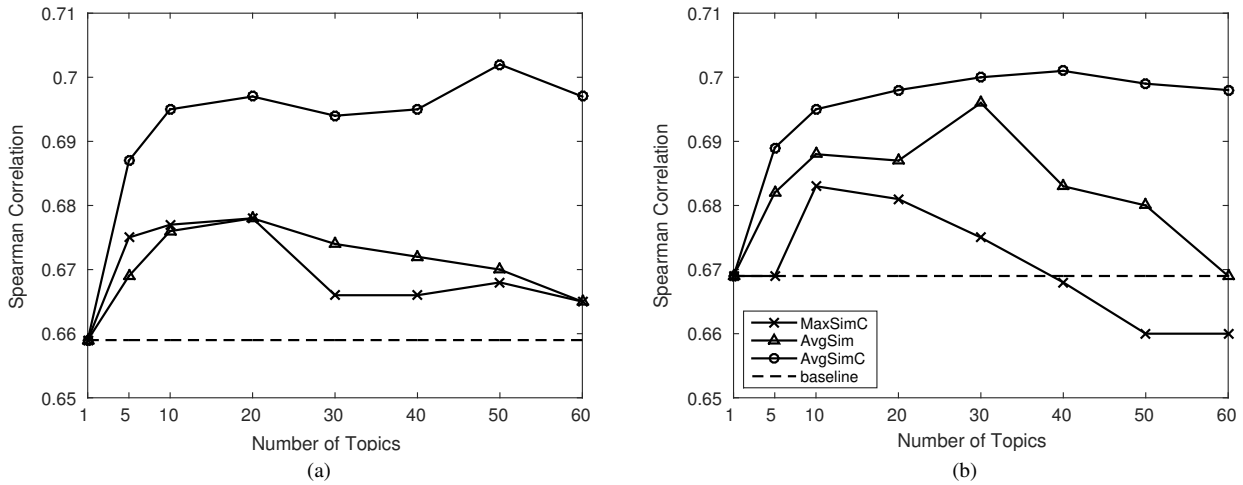

Fig. 3. Spearman ρ correlation for SCWS dataset using the Topic DSMs mixture model (TDSMs) as a function of the number of topics: (a) Web corpus, (b) Wikipedia corpus. For MaxSimC metric (2) only the topic with the maximum posterior probability is considered, for AvgSim (1) all topics contribute equally to the semantic similarity computation and for AvgSimC (3) each candidate topic is weighted by the corresponding posterior probability. The baseline corresponds to a model with a single topic.

MEN dataset, the proposed approach is shown to outperform the baseline for all number of topics for both corpora. The top correlation score (0.838) is achieved for 40 topics using the Web corpus. For the WS-353 dataset, the same combination of topics and corpus provides the top performance (0.727). Overall, the Web corpus appears to yield higher performance compared to the Wikipedia corpus. For larger number of topics (up to 100 – although not shown here) we observed correlation scores comparable to the performance at 60 topics for all datasets and corpora combinations.

The performance of the TDSMs model, for each semantic similarity metric, is depicted in Fig. 3a and 3b as a function of the number of topics for the SCWS dataset. For both corpora, the top performance (0.702 and 0.701) is achieved by the AvgSimC metric when utilizing $40 - 50$ topics.

2) Sentiment Classification: Table III reports the results of the semantic-affective model on the SemEval 2007 Task 14 dataset. It is observed that the top performance is achieved by the use of 30 topics for all three fusion schemes exceeding the baseline (one topic). Specifically, the highest correlation score (0.650) is achieved by the Weighted Fusion scheme.

VI. Discussion

The improvement over the baseline performance, achieved by the proposed approach for the semantic similarity computation task, was clearly demonstrated through the use of three datasets and corpora. We observed that the top performance (0.727), achieved for the WS-353 dataset, is lower compared to the highest correlation score (0.838) obtained for MEN. This

TABLE III
SPEARMAN ρ CORRELATION FOR SENTENCE AFFECTIVE SCORE ESTIMATION ON THE SEMEVAL 2007 TASK 14 DATASET.

Number of Topics	Linear Fusion	Weighted Fusion	Max Fusion
1	0.614	0.627	0.543
10	0.637	0.595	0.563
20	0.626	0.639	0.572
30	**0.646**	**0.650**	**0.603**
40	0.614	0.617	0.551
50	0.641	0.634	0.586
60	0.605	0.608	0.544

can be attributed to the different dataset designs, e.g., the type of semantic relationship, as well as the procedure followed for the collection of human ratings. A critical review of such factors is provided in [47]. Overall, the reported improvement for the MEN dataset is statistically more significant compared to the case of WS-353 due to the larger size of the MEN dataset (i.e., 3000 vs. 353 pairs).

The superiority of the Web corpus, compared to the Wikipedia corpus, for the task of out-of-context semantic similarity computation, can be explained by the underlying corpus creation process. The collection of web document snippets (i.e., a minimum number of 1000) yielded a corpus that deviates from the typical distribution of word frequencies (Zipf law). Specifically, in terms of word frequency statistics, what differentiates the Web corpus from traditional corpora is that words included in the corpus have a minimum number of 1000 occurrences. This applies even for the rarest words included in the lexicon used for the creation of web search queries (for details see [40]). As a result, the respective DSMs encode a wide spectrum of word senses ranging from highly frequent to less frequent word senses. This characteristic yields very good performance of the out-of-context similarity task, where the similarity estimation is not conditioned on specific contexts (senses).

The number of topics constitutes a key parameter of the proposed approach. The identified topics are used for corpus filtering (i.e., creation of sub-corpora) upon which the creation of DSMs is based. In this framework, when computing the similarity between a word pair, we argue that the exploited sub-corpus exhibits two properties: i) The sub-corpus should be semantically coherent, i.e., the two words should appear with their closest word senses, and ii) adequate data should exist enabling the computation of DSMs. Typically, a larger number of topics improves the semantic coherence of the respective sub-corpus (increased topic specificity), but it may cause the fragmentation of the training data lowering the quality of the semantic models.

In order to overcome this issue, the linear regression approach (6) is suitable for selecting the best similarities from the respective topic-based DSMs, for pairs without context. The method surpasses the baseline for very small number of topics but seems to work better for a larger number of topics despite the data fragmentation. This behavior can be explained by considering that without a given context, a word could have an arbitrary number of senses. An augmented sense-space enables to estimate more accurately the general similarity of a pair as a linear combination of different sense-related similarities.

The fusion model (5) provided the best results when all topic groups were used (5 to 100 topics). This is expected as it resembles the functionality of a hierarchical topic model. Hierarchical topic models relax the hypothesis of a single distribution over a corpus. By selecting the maximum similarity over different possible distributions, the actual number of senses assigned to each word can be approached.

Regarding sentiment classification, the weighted fusion scheme (9) provided the best results, as more strongly affective words influence the overall sentiment of the sentence. The experimental findings suggest that 30 is the total number of senses that can be found in the dataset. The two-step process for calculating topic-adapted similarities ensures that for a given sentence, the most semantically relevant topic-DSMs will be used to estimate the representations of its words. This is achieved by taking the thematic domain under which the sentence belongs into consideration using the LDA algorithm.

VII. CONCLUSIONS

In this work, a mixture model of topic-based DSMs was proposed for the computation of semantic similarity between words and the estimation of affective scores for words and sentences. The proposed mixture model was evaluated on out-of-context and in-context datasets. It was shown to outperform the baseline (single topic) model. The good performance of the mixture model can be attributed to the creation of sub-corpora where the words of interest appear with topic-related senses. Furthermore, we incorporated the proposed mixture model into an affective model for estimating sentence-level affective scores. This improved the baseline by 4%.

Future work includes the automatic estimation of the optimal number of topics using semantically-driven criteria. Also, we aim to investigate the normalization and fusion of generic and topic-specific word embeddings according to the polysemy degree of the words subjected to similarity computation. Finally, we intent to validate the universality of the proposed model by experimenting with corpora and evaluation datasets in languages other than English.

ACKNOWLEDGMENTS

This work has been partially funded by the BabyRobot project, supported by the EU Horizon 2020 Program under grant #687831.

REFERENCES

[1] J. Mitchell and M. Lapata, "Composition in distributional models of semantics," *Cognitive science*, vol. 34, no. 8, pp. 1388–1429, 2010.

[2] E. Agirre, M. Diab, D. Cer, and A. Gonzalez-Agirre, "Semeval-2012 task 6: A pilot on semantic textual similarity," in *Proc. International Workshop on Semantic Evaluation (SemEval)*, 2012, pp. 385–393.

[3] P. D. Turney, "Domain and function: A dual-space model of semantic relations and compositions," *Journal of Artificial Intelligence Research*, vol. 44, pp. 533–585, 2012.

[4] S. Padó and M. Lapata, "Dependency-based construction of semantic space models," *Computational Linguistics*, vol. 33, no. 2, pp. 161–199, 2007.

[5] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Computer Research Repository*, 2013.

[7] K. Erk and S. Padó, "Exemplar-based models for word meaning in context," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010, pp. 92–97.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for Information Science*, vol. 41, no. 6, p. 391, 1990.

[10] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50–57.

[11] D. M. Blei and J. D. Lafferty, "A Correlated Topic Model of Science," *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 17–35, 2007.

[12] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *Proc. International Conference on Machine Learning (ICML)*, 2006, pp. 577–584.

[13] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, "Topical word embeddings," in *Proc. AAAI Conference on Artificial Intelligence*, 2015, pp. 2418–2424.

[14] P. Liu, X. Qiu, and X. Huang, "Learning context-sensitive word embeddings with neural tensor skip-gram model," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2015, pp. 1284–1290.

[15] B. Xiang, L. Zhou, and T. Reuters, "Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014, pp. 434–439.

[16] J. Reisinger and R. Mooney, "Mixture Model with Sharing for Lexical Semantics," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2010, pp. 1173–1182.

[17] ——, "Multi-prototype vector-space models of word meaning," in *Proc. Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, 2010, pp. 109–117.

[18] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, "Improving word representations via global context and multiple word prototypes," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012, pp. 873–882.

[19] A. Neelakantan, J. Shankar, A. Passos, and A. McCallum, "Efficient non-parametric estimation of multiple embeddings per word in vector space," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1059–1069.

[20] F. Tian, H. Dai, J. Bian, B. Gao, R. Zhang, E. Chen, and T.-Y. Liu, "A probabilistic model for learning multi-prototype word embeddings," in *Proc. International Conference on Computational Linguistics (COLING)*, 2014, pp. 151–160.

[21] X. Chen, Z. Liu, and M. Sun, "A unified model for word sense representation and disambiguation," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1025–1035.

[22] J. Guo, W. Che, H. Wang, and T. Liu, "Learning sense-specific word embeddings by exploiting bilingual resources," in *Proc. International Conference on Computational Linguistics (COLING)*, 2014, pp. 497–507.

[23] X. Chen, X. Qiu, J. Jiang, and X. Huang, "Gaussian mixture embeddings for multiple word prototypes," *arXiv preprint arXiv:1511.06246*, 2015.

[24] I. Iacobacci, M. T. Pilehvar, and R. Navigli, "Sensembed: Learning sense embeddings for word and relational similarity," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015, pp. 95–105.

[25] S. K. Jauhar, C. Dyer, and E. H. Hovy, "Ontologically grounded multi-sense representation learning for semantic vector space models." in *Proc. Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, 2015, pp. 683–693.

[26] S. Rothe and H. Schütze, "Autoextend: Extending word embeddings to embeddings for synsets and lexemes," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015, pp. 1793–1803.

[27] Z. Wu and C. L. Giles, "Sense-aware semantic analysis: A multi-prototype word representation model using wikipedia," in *Proc. AAAI Conference on Artificial Intelligence*, 2015, pp. 2188–2194.

[28] J. Li and D. Jurafsky, "Do multi-sense embeddings improve natural language understanding?" in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 1722–1732.

[29] H. Amiri, P. Resnik, J. Boyd-Graber, and H. D. III, "Learning text pair similarity with context-sensitive autoencoders," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 1, 2016, pp. 1882–1892.

[30] X. Zheng, J. Feng, Y. Chen, H. Peng, and W. Zhang, "Learning context-specific word/character embeddings," in *Proc. AAAI Conference on Artificial Intelligence*, 2017, pp. 3393–3399.

[31] M. T. Pilehvar and N. Collier, "De-conflated semantic representations," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016, pp. 1680–1690.

[32] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: modeling facets and opinions in weblogs," in *Proc. International Conference on World Wide Web (ICWWW)*, 2007, pp. 171–180.

[33] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proc. ACM Conference on Information and knowledge management*, 2009, pp. 375–384.

[34] Y. Jo and A. H. Oh, "Aspect and sentiment unification model for online review analysis," in *Proc. of the fourth ACM International Conference on Web search and data mining*, 2011, pp. 815–824.

[35] Y. Rao, Q. Li, X. Mao, and L. Wenyin, "Sentiment topic models for social emotion mining," *Information Sciences*, vol. 266, pp. 90–100, 2014.

[36] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, pp. 448–453.

[37] E. Agirre and P. Edmonds, *Word sense disambiguation: Algorithms and applications*. Springer, 2007.

[38] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Emotiword: Affective lexicon creation with application to interaction and multimedia data," in *Proc. International Workshop on computational Intelligence for Multimedia Understanding*, 2011, pp. 30–41.

[39] F. J. Pelletier, "The principle of semantic compositionality," *Topoi*, vol. 13, no. 1, pp. 11–24, 1994.

[40] E. Iosif and A. Potamianos, "Similarity computation using semantic networks created from web-harvested data," *Natural Language Engineering*, vol. 21, no. 1, pp. 49–79, 2015.

[41] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proc. Language Resources and Evaluation Conference (LREC) Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.

[42] E. Bruni, N. Tran, and M. Baroni, "Multimodal Distributional Semantics." *Journal of Artificial Intelligence Resources (JAIR)*, vol. 49, no. 1-47, 2014.

[43] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," in *Proc. International Conference on World Wide Web (ICWWW)*, 2001, pp. 406–414.

[44] M. Bradley and P. Lang, "Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings," The Center for Research in Psychophysiology, University of Florida, Technical report C-1, 1999.

[45] N. Malandrakis, A. Potamianos, K. J. Hsu, K. N. Babeva, M. C. Feng, G. C. Davison, and S. Narayanan, "Affective language model adaptation via corpus selection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4838–4842.

[46] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proc. International Workshop on Semantic Evaluations*, 2007, pp. 70–74.

[47] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *Computational Linguistics*, 2015.