Learning Distributed Representations of Sentences from Unlabelled Data

Felix Hill
Computer Laboratory
University of Cambridge
felix.hill@cl.cam.ac.uk

Kyunghyun Cho
Courant Institute of
Mathematical Sciences
& Centre for Data Science
New York University
kyunghyun.cho@nyu.edu

Anna Korhonen
Department of Theoretical
& Applied Linguistics
University of Cambridge
alk23@cam.ac.uk

Abstract

Unsupervised methods for learning distributed representations of words are ubiquitous in today's NLP research, but far less is known about the best ways to learn distributed phrase or sentence representations from unlabelled data. This paper is a systematic comparison of models that learn such representations. We find that the optimal approach depends critically on the intended application. Deeper, more complex models are preferable for representations to be used in supervised systems, but shallow log-bilinear models work best for building representation spaces that can be decoded with simple spatial distance metrics. We also propose two new unsupervised representation-learning objectives designed to optimise the trade-off between training time, domain portability and performance.

1 Introduction

Distributed representations - dense real-valued vectors that encode the semantics of linguistic units - are ubiquitous in today's NLP research. For single-words or word-like entities, there are established ways to acquire such representations from naturally occurring (unlabelled) training data based on comparatively task-agnostic objectives (such as predicting adjacent words). These methods are well understood empirically (Baroni et al., 2014b) and theoretically (Levy and Goldberg, 2014). The best word representation spaces reflect consistently-observed aspects of human conceptual organisation (Hill et al., 2015b), and can be added as features to improve the performance of numerous language processing systems (Collobert et al., 2011).

By contrast, there is comparatively little consensus on the best ways to learn distributed representations of phrases or sentences.[1] With the advent of deeper language processing techniques, it is relatively common for models to represent phrases or sentences as continuous-valued vectors. Examples include machine translation (Sutskever et al., 2014), image captioning (Mao et al., 2015) and dialogue systems (Serban et al., 2015). While it has been observed informally that the internal sentence representations of such models can reflect semantic intuitions (Cho et al., 2014), it is not known which architectures or objectives yield the 'best' or most useful representations. Resolving this question could ultimately have a significant impact on language processing systems. Indeed, it is phrases and sentences, rather than individual words, that encode the human-like general world knowledge (or 'common sense') (Norman, 1972) that is a critical missing part of most current language understanding systems.

We address this issue with a systematic comparison of cutting-edge methods for learning distributed representations of sentences. We focus on methods that do not require labelled data gathered for the purpose of training models, since such methods are more cost-effective and applicable across languages and domains. We also propose two new phrase or sentence representation learning objectives - *Sequential Denoising Autoencoders* (SDAEs)

[1] See the contrasting conclusions in (Mitchell and Lapata, 2008; Clark and Pulman, 2007; Baroni et al., 2014a; Milajevs et al., 2014) among others.

Proceedings of NAACL-HLT 2016, pages 1367–1377,
San Diego, California, June 12-17, 2016.

and *FastSent*, a sentence-level log-bilinear bag-of-words model. We compare all methods on two types of task - *supervised* and *unsupervised evaluations* - reflecting different ways in which representations are ultimately to be used. In the former setting, a classifier or regression model is applied to representations and trained with task-specific labelled data, while in the latter, representation spaces are directly queried using cosine distance.

We observe notable differences in approaches depending on the nature of the evaluation metric. In particular, deeper or more complex models (which require greater time and resources to train) generally perform best in the supervised setting, whereas shallow log-bilinear models work best on unsupervised benchmarks. Specifically, SkipThought Vectors (Kiros et al., 2015) perform best on the majority of supervised evaluations, but SDAEs are the top performer on paraphrase identification. In contrast, on the (unsupervised) SICK sentence relatedness benchmark, FastSent, a simple, log-bilinear variant of the SkipThought objective, performs better than all other models. Interestingly, the method that exhibits strongest performance across both supervised and unsupervised benchmarks is a bag-of-words model trained to compose word embeddings using dictionary definitions (Hill et al., 2015a). Taken together, these findings constitute valuable guidelines for the application of phrasal or sentential representation-learning to language understanding systems.

2 Distributed Sentence Representations

To constrain the analysis, we compare neural language models that compute sentence representations from unlabelled, naturally-ocurring data, as with the predominant methods for word representations.[2] Likewise, we do not focus on 'bottom up' models where phrase or sentence representations are built from fixed mathe proposed bymatical operations on word vectors (although we do consider a canonical case - see CBOW below); these were already compared by Milajevs et al. (2014). Most space is devoted to our novel approaches, and we refer the reader to the original papers for more details of existing models.

2.1 Existing Models Trained on Text

SkipThought Vectors For consecutive sentences S_{i-1}, S_i, S_{i+1} in some document, the **SkipThought** model (Kiros et al., 2015) is trained to predict target sentences S_{i-1} and S_{i+1} given source sentence S_i. As with all *sequence-to-sequence* models, in training the source sentence is 'encoded' by a Recurrent Neural Network (RNN) (with Gated Recurrent uUnits (Cho et al., 2014)) and then 'decoded' into the two target sentences in turn. Importantly, because RNNs employ a single set of update weights at each time-step, both the encoder and decoder are sensitive to the order of words in the source sentence.

For each position in a target sentence S_t, the decoder computes a softmax distribution over the model's vocabulary. The cost of a training example is the sum of the negative log-likelihood of each correct word in the target sentences S_{i-1} and S_{i+1}. This cost is backpropagated to train the encoder (and decoder), which, when trained, can map sequences of words to a single vector.

ParagraphVector Le and Mikolov (2014) proposed two log-bilinear models of sentence representation. The **DBOW** model learns a vector $\mathbf{s}$ for every sentence S in the training corpus which, together with word embeddings v_w, define a softmax distribution optimised to predict words $w \in S$ given S. The v_w are shared across all sentences in the corpus. In the **DM** model, k-grams of consecutive words $\{w_i \dots w_{i+k} \in S\}$ are selected and $\mathbf{s}$ is combined with $\{v_{w_i} \dots v_{w_{i+k}}\}$ to make a softmax prediction (parameterised by additional weights) of w_{i+k+1}.

We used the Gensim implementation,[3] treating each sentence in the training data as a 'paragraph' as suggested by the authors. During training, both DM and DBOW models store representations for every sentence (as well as word) in the training corpus. Even on large servers it was therefore only possible to train models with representation size 200, and DM models whose combination operation was averaging (rather than concatenation). Unlike other models considered in this section, for both ParagraphVector architectures an inference step is re-

[2] This excludes innovative supervised sentence-level architectures including (Socher et al., 2011; Kalchbrenner et al., 2014) and many others.

[3] https://radimrehurek.com/gensim/

quired after training to estimate sentence representations $\mathbf{s}$ for arbitrary sentences based on the v_w. This additional computation is reflected in the higher encoding time in Table 1 (TE).

Bottom-Up Methods We train **CBOW** and **Skip-Gram** word embeddings (Mikolov et al., 2013b) on the same text corpus as the SkipThought and ParagraphVector models, and compose by elementwise addition as per Mitchell and Lapata (2010).[4]

We also compare to **C-PHRASE** (Pham et al., 2015), an approach that exploits a (supervised) parser to infer distributed semantic representations based on a syntactic parse of sentences. C-PHRASE achieves state-of-the-art results for distributed representations on several evaluations used in this study.[5]

Non-Distributed Baseline We implement a **TFIDF BOW** model in which the representation of sentence S encodes the count in S of a set of feature-words weighted by their *tfidf* in C, the corpus. The feature-words are the 200,000 most common words in C.

2.2 Models Trained on Structured Resources

The following models rely on (freely-available) data that has more structure than raw text.

DictRep Hill et al. (2015a) trained neural language models to map dictionary definitions to pre-trained word embeddings of the words defined by those definitions. They experimented with **BOW** and **RNN** (with LSTM) encoding architectures and variants in which the input word embeddings were either learned or pre-trained (**+embs.**) to match the target word embeddings. We implement their models using the available code and training data.[6]

CaptionRep Using the same overall architecture, we trained (**BOW** and **RNN**) models to map captions in the COCO dataset (Chen et al., 2015) to pre-trained vector representations of images. The image representations were encoded by a deep convolutional network (Szegedy et al., 2014) trained on the ILSVRC 2014 object recognition task (Russakovsky et al., 2014). Multi-modal distributed representations can be encoded by feeding test sentences forward through the trained model.

NMT We consider the sentence representations learned by neural MT models. These models have identical architecture to SkipThought, but are trained on sentence-aligned translated texts. We used a standard architecture (Cho et al., 2014) on all available **En-Fr** and **En-De** data from the 2015 Workshop on Statistical MT (WMT).[7]

2.3 Novel Text-Based Models

We introduce two new approaches designed to address certain limitations with the existing models.

Sequential (Denoising) Autoencoders The SkipThought objective requires training text with a coherent inter-sentence narrative, making it problematic to port to domains such as social media or artificial language generated from symbolic knowledge. To avoid this restriction, we experiment with a representation-learning objective based on *denoising autoencoders* (DAEs). In a DAE, high-dimensional input data is corrupted according to some noise function, and the model is trained to recover the original data from the corrupted version. As a result of this process, DAEs learn to represent the data in terms of features that explain its important factors of variation (Vincent et al., 2008). Transforming data into DAE representations (as a 'pre-training' or initialisation step) gives more robust (supervised) classification performance in deep feedforward networks (Vincent et al., 2010).

The original DAEs were feedforward nets applied to (image) data of fixed size. Here, we adapt the approach to variable-length sentences by means of a noise function $N(S|p_o, p_x)$, determined by free parameters $p_o, p_x \in [0, 1]$. First, for each word w in S, N deletes w with (independent) probability p_o. Then, for each non-overlapping bigram $w_i w_{i+1}$ in S, N swaps w_i and w_{i+1} with probability p_x. We then train the same LSTM-based encoder-decoder architecture as NMT, but with the denoising objective to predict (as target) the original source sentence S given a corrupted version $N(S|p_o, p_x)$ (as source).

[4] We also tried multiplication but this gave very poor results.

[5] Since code for C-PHRASE is not publicly-available we use the available pre-trained model (`http://clic.cimec.unitn.it/composes/cphrase-vectors.html`). Note this model is trained on 3× more text than others in this study.

[6] `https://www.cl.cam.ac.uk/~fh295/`. Definitions from the training data matching those in the WordNet STS 2014 evaluation (used in this study) were excluded.

[7] `www.statmt.org/wmt15/translation-task.html`

The trained model can then encode novel word sequences into distributed representations. We call this model the *Sequential Denoising Autoencoder* (**SDAE**). Note that, unlike SkipThought, SDAEs can be trained on sets of sentences in arbitrary order.

We label the case with no noise (i.e. $p_o = p_x = 0$ and $N \equiv id$) **SAE**. This setting matches the method applied to text classification tasks by Dai and Le (2015). The 'word dropout' effect when $p_o \geq 0$ has also been used as a regulariser for deep nets in supervised language tasks (Iyyer et al., 2015), and for large p_x the objective is similar to word-level 'debagging' (Sutskever et al., 2011). For the SDAE, we tuned p_o, p_x on the validation set (see Section 3.2).[8] We also tried a variant (**+embs**) in which words are represented by (fixed) pre-trained embeddings.

FastSent The performance of SkipThought vectors shows that rich sentence semantics can be inferred from the content of adjacent sentences. The model could be said to exploit a type of *sentence-level Distributional Hypothesis* (Harris, 1954; Polajnar et al., 2015). Nevertheless, like many deep neural language models, SkipThought is very slow to train (see Table 1). FastSent is a simple additive (log-bilinear) sentence model designed to exploit the same signal, but at much lower computational expense. Given a BOW representation of some sentence in context, the model simply predicts adjacent sentences (also represented as BOW) .

More formally, FastSent learns a source u_w and target v_w embedding for each word in the model vocabulary. For a training example S_{i-1}, S_i, S_{i+1} of consecutive sentences, S_i is represented as the sum of its source embeddings $\mathbf{s_i} = \sum_{w \in S_i} u_w$. The cost of the example is then simply:

$$\sum_{w \in S_{i-1} \cup S_{i+1}} \phi(\mathbf{s_i}, v_w) \quad (1)$$

where $\phi(v_1, v_2)$ is the softmax function.

We also experiment with a variant (**+AE**) in which the encoded (source) representation must predict its own words as target in addition to those of adjacent sentences. Thus in FastSent+AE, (1) becomes

$$\sum_{w \in S_{i-1} \cup S_i \cup S_{i+1}} \phi(\mathbf{s_i}, v_w). \quad (2)$$

[8] We searched $p_o, p_x \in \{0.1, 0.2, 0.3\}$ and observed best results with $p_o = p_x = 0.1$.

	OS	R	WO	SD	WD	TR	TE
S(D)AE			✓	2400	100	72*	640
ParagraphVec				100	100	4	1130
CBOW				500	500	2	145
SkipThought	✓		✓	4800	620	336*	890
FastSent	✓			100	100	2	140
DictRep		✓	✓	500	256	24*	470
CaptionRep		✓	✓	500	256	24*	470
NMT		✓	✓	2400	512	72*	720

Table 1: Properties of models compared in this study **OS:** requires training corpus of sentences in order. **R:** requires structured resource for training. **WO:** encoder sensitive to word order. **SD:** dimension of sentence representation. **WD:** dimension of word representation. **TR:** approximate training time (hours) on the dataset in this paper. * indicates trained on GPU. **TE:** approximate time (s) taken to encode 0.5m sentences.

At test time the trained model (very quickly) encodes unseen word sequences into distributed representations with $\mathbf{s} = \sum_{w \in S} u_w$.

2.4 Training and Model Selection

Unless stated above, all models were trained on the Toronto Books Corpus,[9] which has the intersentential coherence required for SkipThought and FastSent. The corpus consists of 70m ordered sentences from over 7,000 books.

Specifications of the models are shown in Table 1. The log-bilinear models (SkipGram, CBOW, ParagraphVec and FastSent) were trained for one epoch on one CPU core. The representation dimension d for these models was found after tuning $d \in \{100, 200, 300, 400, 500\}$ on the validation set.[10] All other models were trained on one GPU. The S(D)AE models were trained for one epoch ($\approx$ 8 days). The SkipThought model was trained for two weeks, covering just under one epoch.[11] For CaptionRep and DictRep, performance was monitored on held-out training data and training was stopped after 24 hours after a plateau in cost. The NMT models were trained for 72 hours.

[9] http://www.cs.toronto.edu/~mbweb/

[10] For ParagraphVec only $d \in \{100, 200\}$ was possible due to the high memory footprint.

[11] Downloaded from https://github.com/ryankiros/skip-thoughts

Dataset		Sentence 1	Sentence 2	/5
	News	*Mexico wishes to guarantee citizens' safety.*	*Mexico wishes to avoid more violence.*	4
	Forum	*The problem is simpler than that.*	*The problem is simple.*	3.8
STS	WordNet	*A social set or clique of friends.*	*An unofficial association of people or groups.*	3.6
2014	Twitter	*Taking Aim #Stopgunviolence #Congress #NRA*	*Obama, Gun Policy and the N.R.A.*	1.6
	Images	*A woman riding a brown horse.*	*A young girl riding a brown horse.*	4.4
	Headlines	*Iranians Vote in Presidential Election.*	*Keita Wins Mali Presidential Election.*	0.4
SICK (test+train)		*A lone biker is jumping in the air.*	*A man is jumping into a full pool.*	1.7

Table 2: Example sentence pairs and 'similarity' ratings from the unsupervised evaluations used in this study.

3 Evaluating Sentence Representations

In previous work, distributed representations of language were evaluated either by measuring the effect of adding representations as features in some classification task - *supervised evaluation* (Collobert et al., 2011; Mikolov et al., 2013a; Kiros et al., 2015) - or by comparing with human relatedness judgements - *unspervised evaluation* (Hill et al., 2015a; Baroni et al., 2014b; Levy et al., 2015). The former setting reflects a scenario in which representations are used to inject general knowledge (sometimes considered as *pre-training*) into a supervised model. The latter pertains to applications in which the sentence representation space is used for direct comparisons, lookup or retrieval. Here, we apply and compare both evaluation paradigms.

3.1 Supervised Evaluations

Representations are applied to 6 sentence classification tasks: paraphrase identification (MSRP) (Dolan et al., 2004), movie review sentiment (MR) (Pang and Lee, 2005), product reviews (CR) (Hu and Liu, 2004), subjectivity classification (SUBJ) (Pang and Lee, 2004), opinion polarity (MPQA) (Wiebe et al., 2005) and question type classification (TREC) (Voorhees, 2002). We follow the procedure (and code) of Kiros et al. (2015): a logistic regression classifier is trained on top of sentence representations, with 10-fold cross-validation used when a train-test split is not pre-defined.

3.2 Unsupervised Evaluations

We also measure how well representation spaces reflect human intuitions of the semantic sentence relatedness, by computing the cosine distance between vectors for the two sentences in each test pair, and correlating these distances with gold-standard human judgements. The SICK dataset (Marelli et al., 2014) consists of 10,000 pairs of sentences and relatedness judgements. The STS 2014 dataset (Agirre et al., 2014) consists of 3,750 pairs and ratings from six linguistic domains. Example ratings are shown in Table 2. All available pairs are used for testing apart from the 500 SICK 'trial' pairs, which are held-out for tuning hyperparameters (representation size of log-bilinear models, and noise parameters in SDAE). The optimal settings on this task are then applied to both supervised and unsupervised evaluations.

4 Results

Performance of the models on the supervised evaluations (grouped according to the data required by their objective) is shown in Table 3. Overall, SkipThought vectors perform best on three of the six evaluations, the BOW DictRep model with pre-trained word embeddings performs best on two, and the SDAE on one. SDAEs perform notably well on the paraphrasing task, going beyond SkipThought by three percentage points and approaching state-of-the-art performance of models designed specifically for the task (Ji and Eisenstein, 2013). SDAE is also consistently better than SAE, which aligns with other findings that adding noise to AEs produces richer representations (Vincent et al., 2008).

Results on the unsupervised evaluations are shown in Table 4. The same DictRep model performs best on four of the six STS categories (and overall) and is joint-top performer on SICK. Of the models trained on raw text, simply adding CBOW word vectors works best on STS. The best performing raw text model on SICK is FastSent, which achieves almost identical performance to C-PHRASE's state-of-the-art performance for a distributed model (Pham et al., 2015). Further, it uses less than a third of the training text and does not

Data	Model	MSRP (Acc / F1)	MR	CR	SUBJ	MPQA	TREC
Unordered Sentences (Toronto Books: 70m sents, 0.9B words)	SAE	74.3 / 81.7	62.6	68.0	86.1	76.8	80.2
	SAE+embs.	70.6 / 77.9	73.2	75.3	89.8	86.2	80.4
	SDAE	**76.4 / 83.4**	67.6	74.0	89.3	81.3	77.6
	SDAE+embs.	73.7 / 80.7	**74.6**	**78.0**	**90.8**	**86.9**	78.4
	ParagraphVec DBOW	72.9 / 81.1	60.2	66.9	76.3	70.7	59.4
	ParagraphVec DM	73.6 / 81.9	61.5	68.6	76.4	78.1	55.8
	Skipgram	69.3 / 77.2	73.6	77.3	89.2	85.0	82.2
	CBOW	67.6 / 76.1	73.6	7730	89.1	85.0	82.2
	Unigram TFIDF	**73.6 / 81.7**	73.7	79.2	90.3	82.4	**85.0**
Ordered Sentences (Toronto Books)	SkipThought	**73.0 / 82.0**	**76.5**	**80.1**	**93.6**	**87.1**	**92.2**
	FastSent	72.2 / 80.3	70.8	78.4	88.7	80.6	76.8
	FastSent+AE	71.2 / 79.1	71.8	76.7	88.8	81.5	80.4
Other structured data resource	NMT En to Fr	69.1 / 77.1	64.7	70.1	84.9	81.5	**82.8**
	NMT En to De	65.2 / 73.3	61.0	67.6	78.2	72.9	81.6
	CaptionRep BOW	73.6 / 81.9	61.9	69.3	77.4	70.8	72.2
	CaptionRep RNN	72.6 / 81.1	55.0	64.9	64.9	71.0	62.4
	DictRep BOW	**73.7 / 81.6**	71.3	75.6	86.6	82.5	73.8
	DictRep BOW+embs.	68.4 / 76.8	**76.7**	**78.7**	**90.7**	**87.2**	81.0
	DictRep RNN	73.2 / 81.6	67.8	72.7	81.4	82.5	75.8
	DictRep RNN+embs.	66.8 / 76.0	72.5	73.5	85.6	85.7	72.0
2.8B words	CPHRASE	72.2 / 79.6	75.7	78.8	91.1	86.2	78.8

Table 3: Performance of sentence representation models on **supervised** evaluations (Section 3.1). Bold numbers indicate best performance in class. Underlined indicates best overall.

require access to (supervised) syntactic representations for training. Together, the results of FastSent on the unsupervised evaluations and SkipThought on the supervised benchmarks provide strong support for the sentence-level distributional hypothesis: the context in which a sentence occurs provides valuable information about its semantics.

Across both unsupervised and supervised evaluations, the BOW DictRep with pre-trained word embeddings exhibits by some margin the most consistent performance. Ths robust performance suggests that DictRep representations may be particularly valuable when the ultimate application is non-specific or unknown, and confirms that dictionary definitions (where available) can be a powerful resource for representation learning.

5 Discussion

Many additional conclusions can be drawn from the results in Tables 3 and 4.

Different objectives yield different representations It may seem obvious, but the results confirm that different learning methods are preferable for different intended applications (and this variation appears greater than for word representations). For instance, it is perhaps unsurprising that SkipThought performs best on TREC because the labels in this dataset are determined by the language immediately following the represented question (i.e. the answer) (Voorhees, 2002). Paraphrase detection, on the other hand, may be better served by a model that focused entirely on the content *within* a sentence, such as SDAEs. Similar variation can be observed in the unsupervised evaluations. For instance, the (multimodal) representations produced by the CaptionRep model do not perform particularly well apart from on the Image category of STS where they beat all other models, demonstrating a clear effect of the well-studied modality differences in representation learning (Bruni et al., 2014).

The nearest neighbours in Table 5 give a more concrete sense of the representation spaces. One notable difference is between (AE-style) models whose semantics come from within-sentence relationships (CBOW, SDAE, DictRep, ParagraphVec) and SkipThought/FastSent, which exploit the context around sentences. In the former case, nearby sentences often have a high proportion of words in common, whereas for the latter it is the general con-

Model	STS 2014							SICK
	News	Forum	WordNet	Twitter	Images	Headlines	All	Test + Train
SAE	17/.16	.12/.12	.30/.23	.28/.22	.49/.46	.13/.11	.12/.13	.32/.31
SAE+embs.	.52/.54	.22/.23	.60/.55	.60/.60	. 64/.64	.41/.41	.42/.43	.47/.49
SDAE	.07/.04	.11/.13	.33/.24	.44/.42	.44/.38	.36/.36	.17/.15	.46/.46
SDAE+embs.	.51/.54	.29/.29	.56/.50	.57/.58	.59/.59	.43/.44	.37/.38	.46/.46
ParagraphVec DBOW	.31/.34	.32/.32	.53/.5	.43/.46	.46/.44	.39/.41	.42/.43	.42/.46
ParagraphVec DM	.42/.46	.33/.34	.51/.48	.54/.57	.32/.30	.46/.47	.44/.44	.44/.46
Skipgram	.56/.59	.42/.42	**.73/.70**	**.71**/.74	.65/.67	**.55**/.58	.62/.63	**.60/.69**
CBOW	**.57/.61**	**.43/.44**	.72/.69	**.71/.75**	.71/.73	**.55/.59**	**.64/.65**	**.60/.69**
Unigram TFIDF	.48/.48	.40/.38	.60/.59	.63/.65	**72/.74**	.49/.49	.58/.57	.52/.58
SkipThought	.44/.45	.14/.15	.39/.34	.42/.43	.55/.60	.43/.44	.27/.29	.57/.60
FastSent	**.58/.59**	**.41**/.36	**.74/.70**	.63/.66	**.74/.78**	.57/.59	**.63/.64**	**.61/.72**
FastSent+AE	.56/ **.59**	**.41/.40**	.69/.64	**.70/.74**	.63/.65	**.58/.60**	.62/.62	.60/.65
NMT En to Fr	.35/.32	.18/.18	.47/.43	.55/.53	.44/.45	.43/.43	.43/.42	.47/.49
NMT En to De	.47/.43	.26/.25	.34/.31	.49/.45	.44/.43	.38/.37	.40/.38	.46/46
CaptionRep BOW	.26/.26	.29/.22	.50/.35	.37/.31	**.78**/.81	.39/.36	.46/.42	.56/.65
CaptionRep RNN	.05/.05	.13/.09	.40/.33	.36/.30	.76/**.82**	.30/.28	.39/.36	.53/.62
DictRep BOW	.62/.67	.42/.40	.81/.81	.62/.66	.66/.68	.53/.58	.62/.65	.57/.66
DictRep BOW+embs.	**.65/.72**	**.49/.47**	**.85/.86**	**.67/.72**	.71/.74	**.57/.61**	**.67/.70**	**.61**/.70
DictRep RNN	.40/.46	.26/.23	.78/.78	.42/.42	.56/.56	.38/.40	.49/.50	.49/.56
DictRep RNN+embs.	.51/.60	.29/.27	.80/.81	.44/.47	.65/.70	.42/.46	.54/.57	.49/.59
CPHRASE	**.69**/.71	.43/.41	.76/.73	.60/.65	.75/.79	**.60/.65**	.65/.67	.60/**.72**

Table 4: Performance of sentence representation models (Spearman/Pearson correlations) on **unsupervised** (relatedness) evaluations (Section 3.2). Models are grouped according to training data as indicated in Table 3.

cepts and/or function of the sentence that is similar, and word overlap is often minimal. Indeed, this may be a more important trait of FastSent than the marginal improvement on the SICK task. Readers can compare the CBOW and FastSent spaces at `http://45.55.60.98/`.

Differences between supervised and unsupervised performance Many of the best performing models on the supervised evaluations do not perform well in the unsupervised setting. In the SkipThought, S(D)AE and NMT models, the cost is computed based on a non-linear decoding of the internal sentence representations, so, as also observed by (Almahairi et al., 2015), the informative geometry of the representation space may not be reflected in a simple cosine distance. The log-bilinear models generally perform better in this unsupervised setting.

Knowledge transfer shows some promise It is notable that, with a few exceptions, the models with pre-trained word embeddings (+embs) outperform those with learned embeddings on both supervised and unsupervised evaluations. In the case of the DictRep models, whose training data is otherwise limited to dictionary definitions, this effect can be considered as a rudimentary form of knowledge transfer. The DictRep+embs model benefits both from the dictionary data and the enhanced lexical semantics acquired from a massive text corpus to build overall higher-quality sentence representations.

Differences in resource requirements As shown in Table 1, different models require different resources to train and use. This can limit their possible applications. For instance, while it was easy to make an online demo for fast querying of near neighbours in the CBOW and FastSent spaces, it was not practical for other models owing to memory footprint, encoding time and representation dimension.

The role of word order is unclear The average scores of models that are sensitive to word order (76.3) and of those that are not (76.6) are approximately the same across supervised evalua-

Query	*If he had a weapon, he could maybe take out their last imp, and then beat up Errol and Vanessa.*	*An annoying buzz started to ring in my ears, becoming louder and louder as my vision began to swim.*
CBOW	*Then Rob and I would duke it out, and every once in a while, he would actually beat me.*	*Louder.*
Skip Thought	*If he could ram them from behind, send them saling over the far side of the levee, he had a chance of stopping them.*	*A weighty pressure landed on my lungs and my vision blurred at the edges, threatening my consciousness altogether.*
FastSent	*Isak's close enough to pick off any one of them, maybe all of them, if he had his rifle and a mind to.*	*The noise grew louder, the quaking increased as the sidewalk beneath my feet began to tremble even more.*
SDAE	*He'd even killed some of the most dangerous criminals in the galaxy, but none of those men had gotten to him like Vitktis.*	*I smile because I'm familiar with the knock, pausing to take a deep breath before dashing down the stairs.*
DictRep (FF+embs.)	*Kevin put a gun to the man's head, but even though he cried, he couldn't tell Kevin anything more.*	*Then gradually I began to hear a ringing in my ears.*
Paragraph Vector (DM)	*I take a deep breath and open the doors.*	*They listened as the motorcycle-like roar of an engine got louder and louder then stopped.*

Table 5: Sample nearest neighbour queries selected from a randomly sampled 0.5m sentences of the Toronto Books Corpus.

Supervised (combined $\alpha = 0.90$)						Unsupervised (combined $\alpha = 0.93$)							
MSRP	MR	CR	SUBJ	MPAQ	TREC	News	Forum	WordNet	Twitter	Images	Headlines	All STS	SICK
0.94 (6)	0.85 (1)	0.86 (4)	0.85 (1)	0.86 (3)	0.89 (5)	0.92 (4)	0.92 (3)	0.92 (4)	0.93 (6)	0.95 (8)	0.92 (2)	0.91 (1)	0.93 (7)

Table 6: Internal consistency (Chronbach's α) among evaluations when individual benchmarks are left out of the (supervised or unsupervised) cohorts. Consistency rank within cohort is in parentheses (1 = most consistent with other evaluations).

tions. Across the unsupervised evaluations, however, BOW models score 0.55 on average compared with 0.42 for RNN-based (order sensitive) models. This seems at odds with the widely held view that word order plays an important role in determining the meaning of English sentences. One possibility is that order-critical sentences that cannot be disambiguated by a robust conceptual semantics (that could be encoded in distributed lexical representations) are in fact relatively rare. However, it is also plausible that current available evaluations do not adequately reflect order-dependent aspects of meaning (see below). This latter conjecture is supported by the comparatively strong performance of TFIDF BOW vectors, in which the effective lexical semantics are limited to simple relative frequencies.

The evaluations have limitations The internal consistency (Chronbach's α) of all evaluations considered together is 0.81 (just above 'acceptable').[12] Table 6 shows that consistency is far higher ('excellent') when considering the supervised or unsupervised tasks as independent cohorts. This indicates that, with respect to common characteristics of sentence representations, the supervised and unsupervised benchmarks do indeed prioritise different properties. It is also interesting that, by this metric, the properties measured by MSRP and image-caption relatedness are the furthest removed from other evaluations in their respective cohorts.

While these consistency scores are a promising sign, they could also be symptomatic of a set of evaluations that are all limited in the same way. The inter-rater agreement is only reported for one of the 8 evaluations considered (MPQA, 0.72 (Wiebe et al., 2005)), and for MR, SUBJ and TREC, each item is only rated by one or two annotators to maximise coverage. Table 2 illustrates why this may be an issue for the unsupervised evaluations; the notion of sentential 'relatedness' seems very subjective. It should be emphasised, however, that the tasks considered in this study are all frequently used for evaluation, and, to our knowledge, there are no existing benchmarks that overcome these limitations.

6 Conclusion

Advances in deep learning algorithms, software and hardware mean that many architectures and objectives for learning distributed sentence representations from unlabelled data are now available to NLP researchers. We have presented the first (to our knowledge) systematic comparison of these methods. We showed notable variation in the performance of approaches across a range of evaluations. Among other conclusions, we found that the op-

[12] wikipedia.org/wiki/Cronbach's_alpha

timal approach depends critically on whether representations will be applied in supervised or unsupervised settings - in the latter case, fast, shallow BOW models can still achieve the best performance. Further, we proposed two new objectives, FastSent and Sequential Denoising Autoencoders, which perform particularly well on specific tasks (MSRP and SICK sentence relatedness respectively).[13] If the application is unknown, however, the best all round choice may be DictRep: learning a mapping of pre-trained word embeddings from the word-phrase signal in dictionary definitions. While we have focused on models using naturally-occurring training data, in future work we will also consider supervised architectures (including convolutional, recursive and character-level models), potentially training them on multiple supervised tasks as an alternative way to induce the 'general knowledge' needed to give language technology the elusive human touch.

Acknowledgments

This work was supported by a Google Faculty Award to AK and FH and a Google European Doctoral Fellowship to FH. Thanks also to Marek Rei, Tamara Polajnar, Laural Rimell, Jamie Ryan Kiros and Piotr Bojanowski for helpful comments.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91.

Amjad Almahairi, Kyle Kastner, Kyunghyun Cho, and Aaron Courville. 2015. Learning distributed representations from reviews for collaborative filtering. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 147–154. ACM.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014a. Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, 9.

[13]We make all code for training and evaluating these new models publicly available, together with pre-trained models and an online demo of the FastSent sentence space.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014b. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res. (JAIR)*, 49:1–47.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of EMNLP*.

Stephen Clark and Stephen Pulman. 2007. Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction*, pages 52–55.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, pages 3061–3069.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350. Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional structure. *Word*.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2015a. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015b. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics.

Yangfeng Ji and Jacob Eisenstein. 2013. Discriminative improvements to distributional sentence similarity. In *EMNLP*, pages 891–896.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of EMNLP*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3276–3284.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of ICML*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan Yulle. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proceedings of ICLR*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*, pages 216–223. Citeseer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Dmitrijs Milajevs, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Matthew Purver. 2014. Evaluating neural word representations in tensor-based compositional settings. In *Proceedings of EMNLP*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL*, pages 236–244.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Donald A Norman. 1972. Memory, knowledge, and the answering of questions.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics.

Nghia The Pham, Germán Kruszewski, Angeliki Lazaridou, and Marco Baroni. 2015. Jointly optimizing word representations for lexical and sentential tasks with the c-phrase model. In *Proceedings of ALC*.

Tamara Polajnar, Laura Rimell, and Stephen Clark. 2015. An exploration of discourse-based sentence spaces for compositional distributional semantics. In *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, page 1.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2014. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Building end-to-end dialogue systems using generative hierarchical neural network models. *arXiv preprint arXiv:1507.04808*.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics.

Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408.

Ellen M Voorhees. 2002. Overview of the trec 2001 question answering track. *NIST special publication*, pages 42–51.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.