Not All Neural Embeddings are Born Equal

Felix Hill
University of Cambridge

KyungHyun Cho
Université de Montréal

Sébastien Jean
Université de Montréal

Coline Devin
Harvey Mudd College

Yoshua Bengio
Université de Montréal, CIFAR Senior Fellow

Abstract

Neural language models learn word representations that capture rich linguistic and conceptual information. Here we investigate the embeddings learned by neural machine translation models. We show that translation-based embeddings outperform those learned by cutting-edge monolingual models at single-language tasks requiring knowledge of conceptual similarity and/or syntactic role. The findings suggest that, while monolingual models learn information about how concepts are related, neural-translation models better capture their true ontological status.

It is well known that word representations can be learned from the distributional patterns in corpora. Originally, such representations were constructed by counting word co-occurrences, so that the features in one word's representation corresponded to other words [11, 17]. Neural language models, an alternative means to learn word representations, use language data to optimise (latent) features with respect to a language modelling objective. The objective can be to predict either the next word given the initial words of a sentence [4, 14, 8], or simply a nearby word given a single cue word [13, 15]. The representations learned by neural models (sometimes called *embeddings*) generally outperform those acquired by co-occurrence counting models when applied to NLP tasks [3].

Despite these clear results, it is not well understood how the architecture of neural models affects the information encoded in their embeddings. Here, we explore this question by considering the embeddings learned by architectures with a very different objective function to monolingual language models: *neural machine translation models*. We show that translation-based embeddings outperform monolingual embeddings on two types of task: those that require knowledge of conceptual similarity (rather than simply association or relatedness), and those that require knowledge of syntactic role. We discuss what the findings indicate about the information content of different embeddings, and suggest how this content might emerge as a consequence of the translation objective.

1 Learning embeddings from language data

Both neural language models and translation models learn real-valued embeddings (of specified dimension) for words in some pre-specified vocabulary, V, covering many or all words in their training corpus. At each training step, a 'score' for the current training example (or batch) is computed based on the embeddings in their current state. This score is compared to the model's objective function, and the error is backpropagated to update both the model weights (affecting how the score is computed from the embeddings) and the embedding features. At the end of this process, the embeddings should encode information that enables the model to optimally satisfy its objective.

1.1 Monolingual models

In the original neural language model [4] and subsequent variants [8], each training example consists of n subsequent words, of which the model is trained to predict the n-th word given the first $n-$

1 words. The model first represents the input as an ordered sequence of embeddings, which it transforms into a single fixed length 'hidden' representation by, e.g., concatenation and non-linear projection. Based on this representation, a probability distribution is computed over the vocabulary, from which the model can sample a guess at the next word. The model weights and embeddings are updated to maximise the probability of correct guesses for all sentences in the training corpus.

More recent work has shown that high quality word embeddings can be learned via models with no nonlinear hidden layer [13, 15]. Given a single word in the corpus, these models simply predict which other words will occur nearby. For each word w in V, a list of training cases $(w, c) : c \in V$ is extracted from the training corpus. For instance, in the *skipgram* approach [13], for each 'cue word' w the 'context words' c are sampled from windows either side of tokens of w in the corpus (with c more likely to be sampled if it occurs closer to w).[1] For each w in V, the model initialises both a cue-embedding, representing the w when it occurs as a cue-word, and a context-embedding, used when w occurs as a context-word. For a cue word w, the model can use the corresponding cue-embedding and all context-embeddings to compute a probability distribution over V that reflects the probability of a word occurring in the context of w. When a training example (w, c) is observed, the model updates both the cue-word embedding of w and the context-word embeddings in order to increase the conditional probability of c.

1.2 Translation-based embeddings

Neural translation models generate an appropriate sentence in their target language S_t given a sentence S_s in their source language [see, e.g., 16, 6]. In doing so, they learn distinct sets of embeddings for the vocabularies V_s and V_t in the source and target languages respectively.

Observing a training case (S_s, S_t), such a model represents S_s as an ordered sequence of embeddings of words from V_s. The sequence for S_s is then encoded into a single representation R_S.[2] Finally, by referencing the embeddings in V_t, R_S and a representation of what has been generated thus far, the model decodes a sentence in the target language word by word. If at any stage the decoded word does not match the corresponding word in the training target S_t, the error is recorded. The weights and embeddings in the model, which together parameterise the encoding and decoding process, are updated based on the accumulated error once the sentence decoding is complete.

Although neural translation models can differ in low-level architecture [7, 2], the translation objective exerts similar pressure on the embeddings in all cases. The source language embeddings must be such that the model can combine them to form single representations for ordered sequences of multiple words (which in turn must enable the decoding process). The target language embeddings must facilitate the process of decoding these representations into correct target-language sentences.

2 Comparing Mono-lingual and Translation-based Embeddings

To learn translation-based embeddings, we trained both the RNN encoder-decoder [*RNNenc*, 7] and the *RNN Search* architectures [2] on a 300m word corpus of English-French sentence pairs. We conducted all experiments with the resulting (English) source embeddings from these models. For comparison, we trained a monolingual skipgram model [13] and its *Glove* variant [15] for the same number of epochs on the English half of the bilingual corpus. We also extracted embeddings from a full-sentence language model [*CW*, 8] trained for several months on a larger 1bn word corpus.

As in previous studies [1, 5, 3], we evaluate embeddings by calculating pairwise (cosine) distances and correlating these distances with (gold-standard) human judgements. Table 1 shows the correlations of different model embeddings with three such gold-standard resources, WordSim-353 [1], MEN [5] and SimLex-999 [10]. Interestingly, translation embeddings perform best on SimLex-999, while the two sets of monolingual embeddings perform better on modelling the MEN and WordSim-353. To interpret these results, it should be noted that SimLex-999 evaluation quantifies conceptual *similarity* (*dog* - *wolf*), whereas MEN and WordSim-353 (despite its name) quantify more general *relatedness* (*dog* - *collar*) [10]. The results seem to indicate that translation-based embeddings better capture similarity, while monolingual embeddings better capture relatedness.

[1] Subsequent variants use different algorithms for selecting the (w, c) from the training corpus [9, 12]

[2] Alternatively, subsequences (phrases) of S_s may be encoded at this stage in place of the whole sentence [2].

		Skipgram	Glove	CW	RNNenc	Search
WordSim-353	ρ	0.52	0.55	0.51	0.57	**0.58**
MEN	ρ	0.44	**0.71**	0.60	0.63	0.62
SimLex-999	ρ	0.29	0.32	0.28	**0.52**	0.49
TOEFL	%	0.75	0.78	0.64	**0.93**	**0.93**
Syn/antonym	%	0.69	0.72	0.75	**0.79**	0.74
teacher	nn	*vocational*	*student*	*student*	*professor*	*instructor*
white	nn	*red*	*red*	*black*	*blank*	*black*
heat	nn	*thermal*	*thermal*	*wind*	*warmth*	*warmth*

Table 1: Translation-based embeddings outperform alternatives on similarity-focused evaluations.

To test this hypothesis further, we ran two more evaluations focused specifically on similarity. The TOEFL synonym test contains 80 cue words, each with four possible answers, of which one is a correct synonym [11]. We computed the proportion of questions answered correctly by each model, where a model's answer was the nearest (cosine) neighbour to the cue word in its vocabulary.[3] In addition, we tested how well different embeddings enabled a supervised classifier to distinguish between synonyms and antonyms. For 500 hand-labelled pairs we presented a Gaussian SVM with the concatenation of the two word embeddings. We evaluated accuracy using 8-fold cross-validation.

As shown in Table 1, translation-based embeddings outperform all monolingual embeddings on these two additional similarity-focused tasks. Qualitative analysis of nearest neighbours (bottom rows) also supports the conclusion that proximity in the translation embedding space corresponds to similarity while proximity in the monolingual embedding space reflects relatedness.

2.1 Quantity of training data

In previous work, monolingual models were trained on corpora many times larger than the English half of our parallel translation corpus. To check if these models simply need more training data to capture similarity as effectively as translation models, we trained them on increasingly large subsets of Wikipedia.[4] The results refute this possibility: the performance of monolingual embeddings on similarity tasks converges well below the level of the translation-based embeddings (Fig. 1).

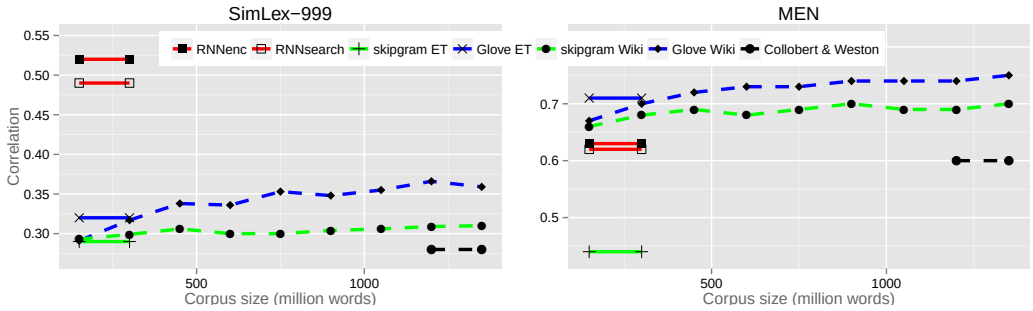

Figure 1: Effect of training corpus size on performance. WordSim-353 results were similar to MEN.

2.2 Analogy questions

Lexical analogy questions are an alternative way of evaluating word representations [13, 15]. In this task, models must identify the correct answer (*girl*) when presented with questions such as '*man* is to *boy* as *woman* is to ...'. For skipgram-style embeddings, it has been shown that if $\mathbf{m}, \mathbf{b}$ and $\mathbf{w}$ are the embeddings for *man*, *boy* and *woman* respectively, the correct answer is often the nearest neighbour in the vocabulary (by cosine distance) to the vector $\mathbf{v} = \mathbf{w} + \mathbf{b} - \mathbf{m}$ [13].

[3] To control for different vocabularies, we restricted the effective vocabulary of each model to the intersection of all model vocabularies, and excluded all questions that contained an answer outiside of this intersection.

[4] We could not do the same for the translation models because of the scarcity of bilingual corpora.

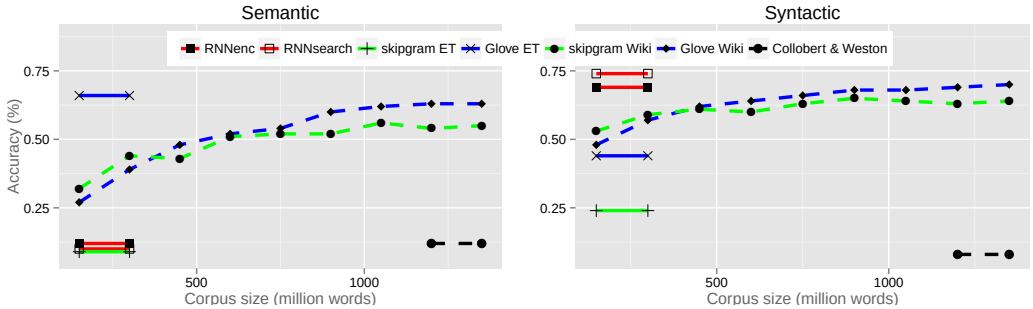

Figure 2: Translation-based embeddings perform best on syntactic analogies (*run,ran: hide, hid*). Monolingual skipgram/Glove models are better at semantic analogies (*father, man; mother, woman*)

We evaluated the embeddings on this task using the same vector-algebra method as [13]. As before we excluded questions containing a word outside the intersection of all model vocabularies, and restricted all answer searches to this reduced vocabulary, leaving 11,166 analogies. Of these, 7219 are classed as 'syntactic', in that they exemplify mappings between parts-of-speech or syntactic roles (*fast, fastest; heavy — heaviest*), and 3947 are classed as 'semantic' (*Ottawa, Canada; Paris — France*), deriving from wider world knowledge. As shown in Fig. 2, the translation-based embeddings seem to yield poor answers to semantic analogy questions, but are very effective for syntactic analogies, outperforming the monolingual embeddings, even those trained on much more data.

3 Conclusions

Neural machine translation models are more effective than monolingual models at learning embeddings that encode information about concept similarity and syntactic role. In contrast, monolingual models encode general inter-concept relatedness (as applicable to semantic analogy questions), but struggle to capture similarity, even when training on larger corpora. For skipgram-style models, whose objective is to predict linguistically collocated pairs, this limitation is perhaps unsurprising, since co-occurring words are, in general, neither semantically nor syntactically similar. However, the fact that it also applies to the full-sentence model *CW* suggests that inferring similarity is problematic for monolingual models even with knowledge of the precise (ordered) contexts of words. This may be because very dissimilar words (such as antonyms) actually often occur in identical linguistic contexts.

When considering the strengths of translation embeddings - similarity and syntactic role - it is notable that each item in the three similarity-focused evaluations consists of word groups or pairs of identical syntactic role. Thus, the strong performance of translation embeddings on similarity tasks cannot be simply a result of their encoding of richer syntactic information. To perform well on SimLex-999, embeddings must encode information approximating what concepts *are* (their function or ontology), even when this contradicts the signal conferred by co-occurrence (as can be the case for related-but-dissimilar concept pairs) [10]. The translation objective seems particularly effective at inducing models to encode such ontological or functional information in word embeddings.

While much remains unknown about this process, one cause might be the different ways in which words partition the meaning space of a language. In cases where a French word has two possible English translations (e.g. *gagner* $\rightarrow$ *win* / *earn*), we note that the (source) embeddings of the two English words are very close. It appears that, since the translation model, which has limited encoding capacity, is trained to map tokens of *win* and *earn* to the same place in the target embedding space, it is efficient to move these concepts closer in the source space. While clear-cut differences in how languages partition meaning space, such as (*gagner* = *win, earn*), may in fact be detrimental to similarity modelling (*win* and *earn* are not synonymous to English speakers), in general, languages partition meaning space in less drastically different ways. We hypothesize that these small differences are the key to how neural translation models approximate ontological similarity so effectively. At the same time, since two dissimilar or even antonymous words in the source language should never correspond to a single word in the target language, these pairs diverge in the embedding space, rendering two antonymous embeddings easily distinguishable from those of two synonyms.

References

[1] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of NAACL-HLT 2009*, 2009.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv:*`1409.0473 [cs.CL]`, September 2014.

[3] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 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, 2014.

[4] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003.

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

[6] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–Decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, October 2014. to appear.

[7] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, October 2014. to appear.

[8] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[9] Felix Hill and Anna Korhonen. Learning abstract concepts from multi-modal data: Since you probably can't see what i mean. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, October 2014.

[10] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*, 2014.

[11] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[12] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, 2014.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[14] Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088, 2009.

[15] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, October 2014.

[16] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[17] Peter D Turney, Patrick Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188, 2010.