EMBEDDING WORD SIMILARITY WITH NEURAL MACHINE TRANSLATION

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

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, or embeddings, that capture rich linguistic and conceptual information. Here we investigate the embeddings learned by *neural machine translation models*, a recently-developed class of neural language model. We show that embeddings from translation models outperform those learned by monolingual models at tasks that require knowledge of both conceptual similarity and lexical-syntactic role. We further show that these effects hold when translating from both English to French and English to German, and argue that the desirable properties of translation embeddings should emerge largely independently of the source and target languages. Finally, we apply a new method for training neural translation models with very large vocabularies, and show that this vocabulary expansion algorithm results in minimal degradation of embedding quality. Our embedding spaces can be queried in an online demo and downloaded from our web page. Overall, our analyses indicate that translation-based embeddings should be used in applications that require concepts to be organised according to similarity and/or lexical function, while monolingual embeddings are better suited to modelling (nonspecific) inter-word relatedness.

1 INTRODUCTION

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 (Landauer & Dumais, 1997; Turney & Pantel, 2010). Neural language models, an alternative method for learning 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 (Bengio et al., 2003; Mnih & Hinton, 2009; Collobert & Weston, 2008), or simply a nearby word given a single cue word (Mikolov et al., 2013; Pennington et al., 2014). The representations learned by neural models (sometimes called *embeddings*) perform very effectively when applied as pre-trained features in a range of NLP applications and tasks (Baroni et al., 2014).

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*. These models have recently emerged as an alternative to statistical, phrase-based translation models, and are beginning to achieve impressive translation performance (Sutskever et al., 2014; Cho et al., 2014a; Bahdanau et al., 2014).

We show that neural translation models are not only a potential new direction for machine translation, but are also an effective means of learning word embeddings. Specifically, translation-based embeddings encode information relating to conceptual similarity (rather than non-specific relatedness or association) and lexical syntactic role more effectively than embeddings from monolingual neural language models. We demonstrate that these properties persist when translating between different language pairs (English-French and English-German). Further, based on the observation of subtle language-specific effects in the embedding spaces, we conjecture as to why similarity dominates over other semantic relations in translation embedding spaces. Finally, we discuss a potential limitation of the application of neural machine translation models for embedding learning - the computational cost of training large vocabularies of embeddings - and show that a novel method for overcoming this issue preserves the quality of translation-based embeddings.

2 Learning Embeddings with Neural Language Models

Both neural language models and neural machine 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 themselves. At the end of this process, the embeddings should encode information that enables the model to optimally satisfy its objective.

2.1 Monolingual models

In the original neural language model (Bengio et al., 2003) and subsequent variants (Collobert & Weston, 2008), training examples consists of an ordered sequence of n words, with the model 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, generally by 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 for 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 simpler models with no nonlinear hidden layer (Mikolov et al., 2013; Pennington et al., 2014). Given a single word or unordered window of words in the corpus, these models predict which 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 according to some algorithm. For instance, in the *skipgram* approach (Mikolov et al., 2013), 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 uses 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] Subsequent variants use different algorithms for selecting the (w, c) from the training corpus (Hill & Korhonen, 2014; Levy & Goldberg, 2014)

2.2 NEURAL MACHINE TRANSLATION MODELS

The objective of neural machine translation models is to generate an appropriate sentence in a target language S_t given a sentence S_s in the source language (see, e.g., Sutskever et al., 2014; Cho et al., 2014a). As a by-product of learning to meet this objective, neural translation models 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), these models represent 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 their low-level architecture (Cho et al., 2014b; Bahdanau et al., 2014), 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.

3 EXPERIMENTS

To learn translation-based embeddings, we trained two different neural machine translation-models. The first is the RNN encoder-decoder (*RNN Enc*, Cho et al., 2014b), which uses a recurrent-neural-network to encode the whole source sentence into a single vector on which the decoding process is conditioned. The second is the *RNN Search* architecture (Bahdanau et al., 2014), which was designed to overcome limitations exhibited by the RNN encoder-decoder when translating very long sentences. RNN Search includes an *attention* mechanism, an additional feed-forward network that learns to attend to different parts of the source sentence when decoding each word in the target sentence. Both models were trained on a 300m word corpus of English-French sentence pairs or a 91m word corpus of English-German sentence pairs, and all experiments were conducted with the resulting (English) source embeddings. For an initial direct comparison, we trained a monolingual skipgram model (Mikolov et al., 2013) and its *Glove* variant (Pennington et al., 2014) for the same number of epochs on the English half of the bilingual corpus. To analyse the effect on embedding quality of increasing the quantity of training data, we also trained these two models on increasingly large random subsamples of Wikipedia text (up to a total of 1.1bn words). Finally, we also extracted embeddings from a full-sentence language model (*CW*, Collobert & Weston, 2008), which was trained for several months on the same Wikipedia 1bn word corpus.

3.1 SIMILARITY AND RELATEDNESS MODELLING

As in previous studies (Agirre et al., 2009; Bruni et al., 2014; Baroni et al., 2014), our initial evaluations involved calculating pairwise (cosine) distances between embeddings and correlating these distances with (gold-standard) human judgements of the strength of relationships between concepts. For this we used three different gold standards: WordSim-353 (Agirre et al., 2009), MEN (Bruni et al., 2014) and SimLex-999 (Hill et al., 2014). Importantly, there is a clear distinction between WordSim-353 and MEN, on the one hand, and SimLex-999, on the other, in terms of the semantic relationship that they quantify. For both WordSim-353 and MEN, annotators were asked to quantify how *related* or *associated* two concepts are. Consequently, pairs such as [*clothes*-*closet*], which are clearly related but ontologically dissimilar, have high ratings in WordSim-353 and MEN. In contrast, such pairs receive a low rating in SimLex-999, where only genuinely *similar* concepts, such as [*coast*- *shore*], receive high ratings.

[2] Alternatively, subsequences (phrases) of S_s may be encoded at this stage in place of the whole sentence (Bahdanau et al., 2014).

			Skipgram	Glove	CW	RNN Enc	RNN Search
Relatedness	WordSim-353	ρ	0.47	0.57	0.51	0.56	**0.60**
	MEN	ρ	0.42	**0.69**	0.60	0.63	0.61
Similarity	SimLex-999	ρ	0.26	0.34	0.28	**0.52**	0.49
	SimLex-Assoc-333	ρ	0.18	0.18	0.07	**0.49**	0.45
	TOEFL	%	0.75	0.78	0.64	**0.93**	0.90
	Syn/antonym	%	0.71	0.73	0.74	**0.79**	0.72

Table 1: Translation-based embeddings (RNN Enc and RNN Search) outperform alternative embedding-learning architectures on tasks that require modelling similarity, but not on tasks that require models to capture relatedness. For fair comparison, all evaluations were restricted to the largest vocabulary common to all models.

	Skipgram	Glove	CW	RNN Enc	RNN Search
teacher	*vocational*	*student*	*student*	*professor*	*instructor*
	in-service	*pupil*	*tutor*	*instructor*	*professor*
	college	*university*	*mentor*	*trainer*	*educator*
eaten	*spoiled*	*cooked*	*baked*	*ate*	*ate*
	squeezed	*eat*	*peeled*	*consumed*	*consumed*
	cooked	*eating*	*cooked*	*tasted*	*eat*
Britain	*Northern*	*Ireland*	*Luxembourg*	*UK*	*England*
	Great	*Kingdom*	*Belgium*	*British*	*UK*
	Ireland	*Great*	*Madrid*	*America*	*Syria*

Table 2: Nearest neighbours (excluding plurals) in the embedding spaces of different model types. All models were trained for 6 epochs on the translation corpus except CW, which was trained for several months on Wikipedia. Translation embedding spaces are oriented according to similarity, whereas embeddings learned by monolingual models are organized according to relatedness.

To reproduce the scores in SimLex-999, models must thus distinguish pairs that are similar from those that are merely related. In particular, this requires models to develop sensitivity to the distinction between synonyms (similar) and antonyms (often strongly related, but highly dissimilar).[3]

Table 1 shows the correlations of neural translation (English-French) embeddings and monolingual embeddings with the three concept-pair-based embedding evaluations. Translation embeddings clearly outperform the monolingual embeddings on SimLex-999, but this clear advantage is not observed on MEN and WordSim-353. Given the aforementioned differences between the evaluations, this suggests that translation-based embeddings better capture similarity, while monolingual embeddings better capture relatedness.

To test this hypothesis further, we ran three more evaluations designed to probe the sensitivity of models to similarity as distinct from general relatedness or association. In the first, we measured performance on SimLex-Assoc-333 (Hill et al., 2014). This evaluation comprises the 333 most related pairs in SimLex-999, according to an independent empirical measure of relatedness (free associate generation **?**). Importantly, the pairs in SimLex-Assoc-333, while all strongly related, still span the full range of similarity scores[4]. Therefore, the extent to which embeddings can model this data reflects their sensitivity to the similarity (or dissimilarity) of two concepts, even in the face of a strong signal in the training data that those concepts are related.

The TOEFL synonym test is another similarity-focused evaluation of embedding spaces. This test contains 80 cue words, each with four possible answers, of which one is a correct synonym (Landauer & Dumais, 1997). 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.[5]. Note that, since TOEFL is a test of synonym recognition, it necessarily requires models to recognise conceptual similarity.

[3]For a more detailed discussion of the similarity/relatedness distinction, see (Hill et al., 2014).

[4]The most dissimilar pair in SimLex-Assoc-333 is [*shrink,grow*] with a score of 0.23. The highest is [*vanish,disappear*] with 9.80.

[5]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 outside of this intersection.

Finally, we tested how well different embeddings enabled a supervised classifier to distinguish between synonyms and antonyms, since synonyms are necessarily similar and people often find antonyms, which are necessarily dissimilar, to be strongly associated. For 744 word pairs hand-selected as either synonyms or antonyms[6], we presented a Gaussian SVM with the concatenation of the two word embeddings. We evaluated accuracy using 10-fold cross-validation.

As shown in Table 1, translation-based embeddings outperform monolingual embeddings on these three additional similarity-focused tasks. The difference is particularly striking on SimLex-Assoc-333, which suggests that the ability to discern similarity from relatedness (when relatedness is high) is perhaps the most clear distinction between the translation embedding spaces and those of monolingual models.

These conslusions are also supported by qualitative analysis of the various embedding spaces. As shown in Table 2, in the translation embedding spaces the nearest neighbours (by cosine distance) to concepts such as *teacher* are genuine synonyms such as *professor* or *instructor*. In contrast, in the monolingual embedding spaces the neighbours of *teacher* include highly related but dissimilar concepts such as *student* or *college*.

3.2 Importance of training data quantity

In previous work, monolingual models were trained on corpora many times larger than the English half of our parallel translation corpus. Indeed, the ability to scale to large quantities of training data was one of the principal motivations behind the skipgram architecture (Mikolov et al., 2013). To check if monolingual models simply need more training data to capture similarity as effectively as translation models, we therefore trained them on increasingly large subsets of Wikipedia.[7] As shown in Figure 1, this is not in fact the case. The performance of monolingual embeddings on similarity tasks remains well below the level of the translation-based embeddings, and appears to have converged before the quantity of training data reaches its maximum.

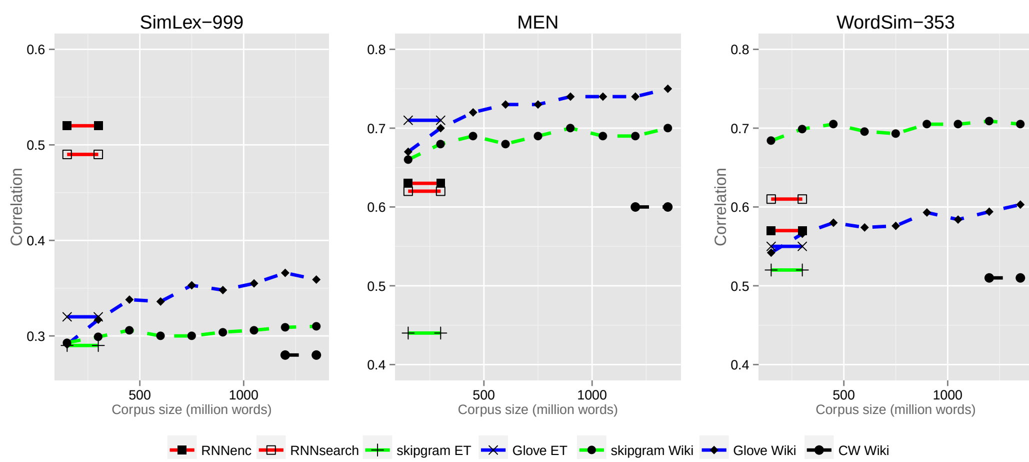

Figure 1: The effect of increasing the amount of training data on the quality of monolingual embeddings, based on similarity-based evaluations (SimLex-999) and two relatedness-based evaluations (MEN and WordSim-353). *ET* in the legend indicates models trained on the English half of the translation corpus. *Wiki* indicates models trained on Wikipedia.

3.3 Analogy resolution

Lexical analogy questions have recently been used as an alternative way of evaluating word representations. In this task, models must identify the correct answer (*girl*) when presented with analogy questions such as ‘*man* is to *boy* as *woman* is to ?’. It has been shown that Skipgram-style models are

[6]Available online at `http://www.cl.cam.ac.uk/~fh295/`.

[7]We did not do the same for our translation models because sentence-aligned bilingual corpora of comparable size do not exist.

surprisingly effective at answering such questions (Mikolov et al., 2013; Pennington et al., 2014). This is because, if $\mathbf{m}$, $\mathbf{b}$ and $\mathbf{w}$ are skigram-style embeddings for *man*, *boy* and *woman* respectively, the correct answer is often the nearest neighbour in the vocabulary (by cosine distance) to the vector $\mathbf{v} = \mathbf{w} + \mathbf{b} - \mathbf{m}$.

We evaluated all embeddings on analogy questions using exactly the same vector-algebra method as Mikolov et al. (2013). As with the analyses in the previous section, for fair comparison we excluded questions containing a word outside the intersection of all model vocabularies, and restricted all answer searches to this reduced vocabulary. This left 11,166 analogies. Of these, 7219 are classed as 'syntactic', in that they exemplify mappings between parts-of-speech or syntactic roles (e.g. *fast* is to *fastest* as *heavy* is to *heaviest*), and 3947 are classed as 'semantic' (*Ottawa* is to *Canada* as *Paris* is to *France*), since successful answering seems to rely on some (world) knowledge of the concepts themselves.

As shown in Fig. 2, translation-based embeddings yield relatively poor answers to semantic analogy questions. It appears that the translation objective prevents the embedding space from developing the same linear, geometric regularities as monolingual (skipgram-style) models with respect to semantic organisation. Indeed, at least in the case of the Glove model, this effect is independent of both the domain and size of the training data, since embeddings from the Glove model trained on only the English half of the translation corpus still clearly outperform the translation embeddings.

Note that in resolving the (semantic) analogy question *A is to B as C is to ?* using the linear vector algebra method, the position in the embedding space of *A* relative to *C*, or *B* relative to the answer, *D*, is unimportant. It is sufficient to ensure that the relative position of A to B is the same as that between C and D. But the relationship between A and B, or between C and D, is almost never one of similarity (whereas *A* and *C*, or *B* and *D*, are indeed in general similar concepts (Turney, 2012)). Thus, embeddings oriented according to similarity, and not more general inter-concept relations, may struggle at the semantic analogy task, and that appears to be borne out by our analyses of both translation-based and skipgram-like embeddings in this case.

Translation-based embeddings, on the other hand, are effective for answering syntactic analogies using the vector algebra method. They perform comparably to or even better than monolingual embeddings when trained on less data (albeit bilingual data). It is perhaps unsurprising that the translation objective incentivises the encoding of a high degree of lexical syntactic information, since coherent target-language sentences could not be generated without knowledge of the parts-of-speech, tense or case of its vocabulary items. It is also notable that items in the similarity-focused evaluations of the previous section (SimLex-999 and TOEFL) consist of word groups or pairs that have identical syntactic role. Thus, even though lexical semantic information can be pertinent to conceptual similarity (Levy & Goldberg, 2014), the lexical syntactic and conceptual properties of translation embeddings are in some sense independent of one another.

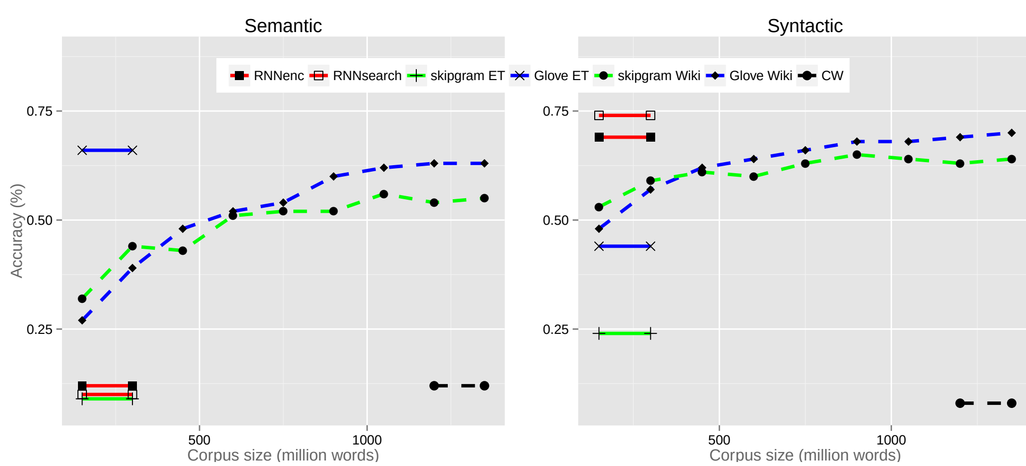

Figure 2: Translation-based embeddings perform best on syntactic analogies (*run,ran: hide, hid*). Monolingual skipgram/Glove models are better at semantic analogies (*father, man; mother, woman*)

4 EFFECT OF TARGET LANGUAGE

To better understand why a translation objective yields embedding spaces with particular properties, we trained the RNN Search architecture to translate from English to German.

		EN-FR	EN-DE
WordSim-353	ρ	0.60	**0.61**
MEN	ρ	0.61	**0.62**
SimLex-999	ρ	0.49	**0.50**
SimLex-Assoc-333	ρ	0.45	**0.47**
TOEFL	%	0.90	**0.93**
Syn/antonym	%	**0.72**	0.70
Syntactic analogies	%	**0.73**	0.62
Semantic analogies	%	0.10	**0.11**

	'earned'	'castle'	'money'
	gained	*chateau*	***silver***
EN-FR	***won***	*palace*	*funds*
	acquired	*fortress*	*cash*
	gained	*chateau*	*funds*
EN-DE	*deserved*	*palace*	*cash*
	accummulated	***padlock***	*resources*

Table 3: Comparison of embeddings learned by RNN Search models translating between English-French (EN-FR) and English-German (EN-DE) on all semantic evaluations (left) and nearest neighbours of selected cue words (right). Bold italics indicate target-language-specific effects. Evaluation items and vocabulary searches were restricted to words common to both models.

As shown in Table 3 (left side), the performance of the source (English) embeddings learned by this model was comparable to that of those learned by the English-to-French model on all evaluations, even though the English-German training corpus (91 million words) was notably smaller than the English-French corpus (300m words). This evidence shows that the desirable properties of translation embeddings highlighted thus far are not particular to English-French translation, and can also emerge when translating to a different language family, with different word ordering conventions.

5 HOW SIMILARITY EMERGES

Although neural translation models appear to encode both conceptual similarity and syntactic information for any source and target languages, it is not the case that embedding spaces will always be identical. Interrogating the nearest neighbours of the source embedding spaces of the English-French and English-German models reveals occasional language-specific effects. As shown in Table 3 (right side), the neighbours for the word *earned* in the English-German model are as one might expect, whereas the neighbours from the English-French model contain the somewhat unlikely candidate *won*. In a similar vein, while the neighbours of the word *castle* from the English-French model are unarguably similar, the neighbours from the English-German model contain the word *padlock*.

These infrequent but striking differences between the English-German and English-French source embedding spaces indicate how similarity might emerge effectively in neural translation models. Tokens of the French verb *gagner* have (at least) two possible English translations (*win* and *earn*). 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. Since *win* and *earn* map directly to two different verbs in German, this effect is not observed. On the other hand, the English nouns *castle* and *padlock* translate to a single noun (*Schloss*) in German, but different nouns in French. Thus, *padlock* and *castle* are only close in the source embeddings from the English-German model.

Based on these considerations, we can conjecture that the following condition on the semantic configuration between two language is crucial to the effective induction of lexical similarity.

(1) For s_1 and s_2 in the source language, there is some t in the target language such that there are sentences in the training data in which s_1 translates to t and sentences in which s_2 translates to t.

if and only if

(2) s_1 and s_2 are semantically similar.

		RNN Search EN-FR	RNN Search EN-DE	RNN Search-LV EN-FR	RNN Search-LV EN-DE
WordSim-353	ρ	0.60	**0.61**	0.59	0.57
MEN	ρ	0.61	**0.62**	**0.62**	0.61
SimLex-999	ρ	0.49	0.50	**0.51**	0.50
SimLex-Assoc-333	ρ	0.45	**0.47**	**0.47**	0.46
TOEFL	%	0.90	0.93	0.93	**0.98**
Syn/antonym	%	0.72	0.70	**0.74**	0.71
Syntactic analogies	%	**0.73**	0.62	0.71	0.62
Semantic analogies	%	0.10	0.11	0.08	**0.13**

Table 4: Comparison of embeddings learned by the original (RNN Search - 30k French words, 50k German words) and extended-vocabulary (RNN Search-LV -500k words) models translating from English to French (EN-FR) and from English to German (EN-DE). For fair comparisons, all evaluations were restricted to intersection of all model vocabularies.

Of course, this condition is not true in general. However, we propose that the extent to which it holds over all possible word pairs corresponds to the quality of similarity induction in the translation embedding space. Note that strong polysemy in the target language, such as *gagner = win, earn*, can lead to cases in which 1 is satisfied but 2 is not. The conjecture claims that these cases are detrimental to the quality of the embedding space (at least with regards to similarity). In practice, qualitative analyses of the embedding spaces and native speaker intuitions suggest that such cases are comparatively rare. Moreover, when such cases are observed, s_1 and s_2, while perhaps not similar, are not strongly dissimilar. This could explain why related but strongly dissimilar concepts (such as antonym pairs) do not converge in the translation embedding space.

6 Overcoming the Vocabulary Size Problem

One potential drawback to using neural machine translation models for learning word embeddings is the computational cost of training such a model on large vocabularies. To generate a target language sentence, the model must repeatedly compute a softmax distribution over the target vocabulary. This computation scales with the size of the vocabulary and must be repeated for each word generated in the output sentence, so that training models with large output vocabularies is challenging. Moreover, while the same computational bottleneck does not apply to the encoding process or source vocabulary, there is no way in which a translation model could learn a high quality source embedding for a word if the plausible translations were outside its vocabulary. Thus, limitations on the size of the target vocabulary effectively limit the scope of neural translation models as representation-learning tools. This contrasts with shallow monolingual models such as skipgram or Glove, which efficiently compute a distribution over a large target vocabulary using either a hierarchical softmax (Morin & Bengio, 2005) or approximate methods such as negative sampling (Mikolov et al., 2013), and thus can learn large vocabularies of both source and target embeddings.

A recently proposed solution to this problem enables neural machine translation models to be trained with larger target vocabularies (and hence larger meaningful source vocabularies) at comparable computational cost to training with a small target vocabulary (Jean et al., 2014). The algorithm uses (biased) importance sampling (Bengio & Sénécal, 2003) to approximate the probability distribution of words over a large target vocabulary with finite set of distributions over subsets of that vocabulary. Despite this element of approximation in the decoder, extending the effective target vocabulary in this way significantly improves translation performance, since the model can make sense of more sentences in the training data and encounters fewer unknown words at test time. In terms of representation learning, the method provides a means to scale up the neural translation approach to vocabularies as large as those learned by monolingual models. However, given that the method replaces an exact calculation with an approximate one, we tested how the quality of source embeddings is affected by scaling up the target language vocabulary in this way.

As shown in Table 4, there is no significant degradation of embedding quality when scaling to large vocabularies with using the approximate decoder. Note that for a fair comparison we filtered these evaluations to only include items that are present in the smaller vocabulary. Thus, the numbers do not directly reflect the quality of the additional 470k embeddings learned by the extended vocabulary models, which one would expect to be lower since they are words of lower frequency. All embeddings can be downloaded from `http://www.cl.cam.ac.uk/~fh295/`, and the embeddings from the smaller vocabulary models can be interrogated at `http://lisa.iro.umontreal.ca/mt-demo/embs/`.

7 Conclusion

In this work, we have shown that the embedding spaces from neural machine translation models are orientated more towards conceptual similarity than those of monolingual models, and that translation embedding spaces also reflect richer lexical syntactic information. To perform well on similarity evaluations such as SimLex-999, embeddings must distinguish information pertinent to what concepts *are* (their function or ontology) from information reflecting other non-specific inter-concept relationships. Concepts that are strongly related but dissimilar (such as antonyms) are particularly challenging in this regard (Hill et al., 2014). We suggested how the nature of the semantic correspondence between the words in languages can endow embeddings from neural translation models with the information needed to reflect human intuitions of similarity, even in these challenging cases.

It should be noted that the RNN encoder-decoder, and neural machine translation models in general, are not the first models to learn word embeddings from bilingual data. Various studies have shown that models trained on aligned bilingual data can embed words from different languages in a 'multilingual' space reflecting cross-lingual word correspondences (A. Klementiev & Bhattarai, 2012; Hermann & Blunsom, 2014; Chandar et al., 2014). Such bilingual embedding models have be applied to tasks involving semantic transfer between languages, including cross-lingual document classification. In contrast to these contributions, our results describe the organisation of word embeddings with respect to other words of the same language (even though they are learned via the translation objective). Interestingly, given our proposal for how it emerges in bilingual models (Section 5), the full encoder-decoder translation pipeline may not be necessary for effective similarity learning. Rather, the objective of equalizing encoded sentence representations from distinct languages, as in Hermann & Blunsom (2014); Chandar et al. (2014), may be sufficient. Of course, the resulting embeddings may not have the same lexical syntactic richness as neural machine translation embeddings, unless the encoder in such models is fully sensitive to word order.

Not all word embeddings learned from running text are born equal. Depending on the application, it may be preferable to use those learned by neural machine translation models rather than those learned by monolingual neural language models. For decades, distributional semantic models have aimed to exploit Firth's famous *distributional hypothesis* to induce word meanings from (monolingual) text. However, the hypothesis also betrays the weakness of the monolingual distributional approach when it comes to learning conceptual representations as rich as those learned by humans. For while it is undeniable that "words which are similar in meaning appear in similar distributional contexts" (Firth, 1957), the converse assertion, which is what really matters, is only sometimes true.

Acknowledgments

The authors would like to thank the developers of Theano Bergstra et al. (2010); Bastien et al. (2012). We acknowledge the support of the following agencies for research funding and computing support: St John's College Cambridge, NSERC, Calcul Québec, Compute Canada, the Canada Research Chairs and CIFAR.

References

A. Klementiev, I. Titov and Bhattarai, B. Inducing crosslingual distributed representations of words. In *COLING*, 2012.

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

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

Baroni, Marco, Dinu, Georgiana, and Kruszewski, Germán. 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.

Bastien, Frédéric, Lamblin, Pascal, Pascanu, Razvan, Bergstra, James, Goodfellow, Ian J., Bergeron, Arnaud, Bouchard, Nicolas, and Bengio, Yoshua. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

Bengio, Yoshua and Sénécal, Jean-Sébastien. Quick training of probabilistic neural nets by importance sampling. In *Proceedings of AISTATS 2003*, 2003.

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

Bergstra, James, Breuleux, Olivier, Bastien, Frédéric, Lamblin, Pascal, Pascanu, Razvan, Desjardins, Guillaume, Turian, Joseph, Warde-Farley, David, and Bengio, Yoshua. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation.

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

Chandar, Sarath, Lauly, Stanislas, Larochelle, Hugo, Khapra, Mitesh M., Ravindran, Balaraman, Raykar, Vikas, and Saha, Amrita. An Autoencoder Approach to Learning Bilingual Word Representations. In *NIPS*, 2014.

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

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

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

Firth, J, R. *A synopsis of linguistic theory 1930-1955*, pp. 1–32. Oxford: Philological Society, 1957.

Hermann, Karl Moritz and Blunsom, Phil. Multilingual Distributed Representations without Word Alignment. In *Proceedings of ICLR*, April 2014. URL http://arxiv.org/abs/1312.6173.

Hill, Felix and Korhonen, Anna. 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.

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

Jean, Sébastien, Cho, Kyunghyun, Memisevic, Roland, and Bengio, Yoshua. On using very large target vocabulary for neural machine translation. *arXiv preprint arXiv:1412.2007*, 2014.

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

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

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

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

Morin, Frederic and Bengio, Yoshua. Hierarchical probabilistic neural network language model. In *AISTATS*, volume 5, pp. 246–252. Citeseer, 2005.

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

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

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

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