Mach Translat (2017) 31:3–18
DOI 10.1007/s10590-017-9194-2

The representational geometry of word meanings acquired by neural machine translation models

Felix Hill[1] · **Kyunghyun Cho**[2] · **Sébastien Jean**[2] · **Yoshua Bengio**[3]

Received: 15 May 2016 / Accepted: 28 March 2017 / Published online: 29 April 2017

Abstract This work is the first comprehensive analysis of the properties of word embeddings learned by neural machine translation (NMT) models trained on bilingual texts. We show the word representations of NMT models outperform those learned from monolingual text by established algorithms such as Skipgram and CBOW on tasks that require knowledge of semantic similarity and/or lexical–syntactic role. These effects hold when translating from English to French and English to German, and we argue that the desirable properties of NMT word embeddings should emerge largely independently of the source and target languages. Further, we apply a recently-proposed heuristic method for training NMT models with very large vocabularies, and show that this vocabulary expansion method results in minimal degradation of embedding quality. This allows us to make a large vocabulary of NMT embeddings available for future research and applications. Overall, our analyses indicate that NMT embeddings should be used in applications that require word concepts to be organised according to similarity and/or lexical function, while monolingual embeddings are better suited to modelling (nonspecific) inter-word relatedness.

Keywords Machine translation · Word embeddings · Representation

✉ Felix Hill
felixhill@google.com

1 Deepmind, 7 Pancras Square, London NC14AG, UK

2 Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

3 MILA, Université de Montréal, Montreal, Canada

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 and Dumais 1997; Turney and 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 and Hinton 2009; Collobert and Weston 2008), or simply a nearby word given a single cue word (Mikolov et al. 2013b; Pennington et al. 2014).

The representations learned by neural language models (sometimes called *embeddings*), are an example of successful and effective unsupervised learning. Word embeddings acquired from raw (unlabelled) text via task-agnostic learning objectives perform very effectively when applied as pre-trained features in a range of NLP applications, including document classification (Kusner et al. 2015), information retrieval (Weston et al. 2010), semantic role labelling (Collobert et al. 2011) and analogy detection (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 contribute to this understanding by considering the embeddings learned by architectures with a very different objective function: *neural machine translation* (NMT) *models*. NMT models have recently emerged as an alternative to statistical, phrase-based translation models, and are beginning to achieve impressive translation performance (Kalchbrenner and Blunsom 2013; Devlin et al. 2014; Sutskever et al. 2014).

In this article, we show that NMT models are not only a potential new direction for machine translation, but are also a means to acquire word embeddings with interesting and useful properties. 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). 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 NMT 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 aforementioned properties of translation-based embeddings.

2 Learning embeddings with neural language models

All neural language models, including NMT models, learn real-valued embeddings 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 and Weston 2008), training examples consist 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 layers (Mikolov et al. 2013b; 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. 2013b), 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.

2.2 Bilingual representation-learning models

Various studies have demonstrated that word representations can also be effectively learned from bilingual corpora, aligned at the document, paragraph or word level (Haghighi et al. 2008; Vulić et al. 2011; Mikolov et al. 2013a; Hermann and Blunsom 2014; Chandar et al. 2014). These approaches aim to represent the words from two (or more) languages in a common vector space so that words in one language are close to words with similar or related meanings in the other. The resulting multilingual embedding spaces have been effectively applied to bilingual lexicon extraction (Haghighi et al. 2008; Vulić et al. 2011; Mikolov et al. 2013a) and document

[1] Subsequent variants use different algorithms for selecting the (w, c) from the training corpus (Hill and Korhonen 2014; Levy and Goldberg 2014).

classification (Klementiev et al. 2012a; Hermann and Blunsom 2014; Chandar et al. 2014; Kočiský et al. 2014).

We focus our analysis on two representatives of this class of (non-NMT) bilingual model. The first is that of (Hermann and Blunsom 2014), whose embeddings improve on the performance of (Klementiev et al. 2012b) in document classification applications. As with the NMT models introduced in the next section, this model can be trained directly on bitexts aligned only at the sentence rather than word level. When training, for aligned sentences S_E and S_F in different languages, the model computes representations R_E and R_F by summing the embeddings of the words in S_E and S_F respectively. The embeddings are then updated to minimise the divergence between R_E and R_F (since they convey a common meaning). A noise-contrastive loss function ensures that the model does not arrive at trivial (e.g. all zero) solutions to this objective. (Hermann and Blunsom 2014) show that, despite the lack of prespecified word alignments, words in the two languages with similar meanings converge in the bilingual embedding space.[2]

The second model we examine is that of (Faruqui and Dyer 2014). Unlike the models described above, (Faruqui and Dyer 2014) showed explicitly that projecting word embeddings from two languages (learned independently) into a common vector space can favourably influence the orientation of word embeddings when considered in their monolingual subspace; i.e., relative to other words in their own language. In contrast to the other models considered in this paper, the approach of (Faruqui and Dyer 2014) requires bilingual data to be aligned at the word level.

2.3 Neural machine translation models

The objective of NMT is to generate an appropriate sentence in a target language S_t given a sentence S_s in the source language (see e.g. Kalchbrenner and Blunsom 2013; Sutskever et al. 2014). As a by-product of learning to meet this objective, NMT 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. 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 NMT models can differ in the details of their architecture (Kalchbrenner and Blunsom 2013; Cho et al. 2014; Bahdanau et al. 2015), the translation objective

[2] The models of Chandar et al. (2014) and Hermann and Blunsom (2014) both aim to minimise the divergence between source and target language sentences represented as sums of word embeddings. Because of these similarities, we do not compare with both in this paper.

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 NMT models. The first is the RNN encoder–decoder (Cho et al. 2014), referred to as *RNNenc*, which uses a recurrent neural network to encode all of the source sentence into a single vector on which the decoding process is conditioned. The second is the *RNN Search* architecture (Bahdanau et al. 2015), which was designed to overcome limitations exhibited by the RNN encoder–decoder when translating very long sentences. RNN Search includes a *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.[3] Both models use *gated recurrent units* (GRUs) (Cho et al. 2014) as activation functions in the recurrent parts of the encoder and decoder, with source and target vocabularies restricted to the 30,000 most frequent words in both languages. The models were trained on a 348m word corpus of English–French sentence pairs or a 91m word corpus of English–German sentence pairs.[4]

To explore the properties of bilingual embeddings learned via objectives other than direct translation, we trained the *BiCVM* model of (Hermann and Blunsom 2014) on the same data, and also downloaded the projected embeddings of (Faruqui and Dyer 2014), *FD*, trained on a bilingual corpus of comparable size ($\approx$ 300 million words per language).[5] Finally, for an initial comparison with monolingual models, we trained a conventional skipgram model (Mikolov et al. 2013b) 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 then trained the monolingual models on increasingly large random subsamples of Wikipedia text (up to a total of 1.1bn words). Lastly, we extracted embeddings from a full-sentence language model (referred to here as *CW* Collobert and Weston 2008), which was trained for several months on the same Wikipedia 1bn word corpus. Note that increasing the volume of training data for the bilingual (and NMT) models was not possible because of the limited size of available sentence-aligned bitexts.

[3] Access to source code and limited GPU time prevent us from training and evaluating the embeddings from other NMT models such as that of Kalchbrenner and Blunsom (2013), Devlin et al. (2014) and Sutskever et al. (2014). The underlying principles of encoding–decoding also apply to these models, and we expect the embeddings would exhibit similar properties to those analysed here.

[4] These corpora were produced from the WMT14 parallel data after conducting the data-selection procedure described by Cho et al. (2014).

[5] Available from http://www.cs.cmu.edu/mfaruqui/soft.html. The available embeddings were trained on English–German aligned data, but the authors report similar results for English–French.

Table 1 NMT embeddings (RNNenc and RNNsearch) clearly outperform alternative embedding-learning architectures on tasks that require modelling similarity (italics), but not on tasks that reflect relatedness

		Monolingual models			Biling. models		NMT models	
		Skipgram	Glove	CW	FD	BiCVM	RNNenc	RNNsearch
WordSim-353	ρ	0.52	0.55	0.51	**0.69**	0.50	0.57	0.58
MEN	ρ	0.44	0.71	0.60	**0.78**	0.45	0.63	0.62
SimLex-999	ρ	*0.29*	*0.32*	*0.28*	*0.39*	*0.36*	**0.52**	*0.49*
SimLex-333	ρ	*0.18*	*0.18*	*0.07*	*0.24*	*0.34*	**0.49**	*0.45*
TOEFL	%	*0.75*	*0.78*	*0.64*	*0.84*	*0.87*	**0.93**	**0.93**
Syn/antonym	%	*0.69*	*0.72*	*0.75*	*0.76*	*0.70*	**0.79**	*0.74*

Bilingual embedding spaces learned without the translation objective are somewhere between these two extremes

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 rate 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.

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).[6]

Table 1 shows the correlations of NMT (English–French) embeddings, other bilingually-trained embeddings and monolingual embeddings with these three lexical gold-standards. NMT outperform monolingual embeddings, and, to a lesser extent, the other bilingually trained embeddings, on SimLex-999. However, this clear advantage is not observed on MEN and WordSim-353, where the projected embeddings of (Faruqui and Dyer 2014), which were tuned for high performance on WordSim-353, perform best. Given the aforementioned differences between the evaluations, this suggests that bilingually-trained embeddings, and NMT based embeddings in particular, better capture similarity, whereas monolingual embedding spaces are orientated more towards relatedness.

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

To test this hypothesis further, we ran three more evaluations designed to probe the sensitivity of models to similarity as distinct from 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 Nelson et al. 2004). Importantly, the pairs in SimLex-Assoc-333, while all strongly related, still span the full range of similarity scores.[7] 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 and 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.[8] Note that, since TOEFL is a test of synonym recognition, it necessarily requires models to recognise similarity as opposed to relatedness.

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[9] we presented a Gaussian SVM with the concatenation of the two word embeddings. We evaluated accuracy using 10-fold cross-validation (Fig. 1).

As shown in Table 1, with these three additional similarity-focused tasks we again the same pattern of results. NMT embeddings outperform other bilingually-trained embeddings which in turn outperform monolingual models. 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 bilingual spaces and those of monolingual models.

These conclusions are also supported by qualitative analysis of the various embedding spaces. As shown in Table 2, in the NMT embedding spaces the nearest neighbours (by cosine distance) to concepts such as *teacher* are genuine synonyms such as *professor* or *instructor*. The bilingual objective also seems to orientate the non-NMT embeddings towards semantic similarity, although some purely related neighbours are also oberved. In contrast, in the monolingual embedding spaces the neighbours of *teacher* include highly related but dissimilar concepts such as *student* or *college*.

It is notable that, while all NMT embedding spaces reflect similarity better than other embedding spaces, those acquired by the RNNenc model achieve better performance than those from the RNNsearch across our evaluations. This may be a simple

[7] 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.

[8] 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.

[9] Available online at http://www.cl.cam.ac.uk/fh295/.

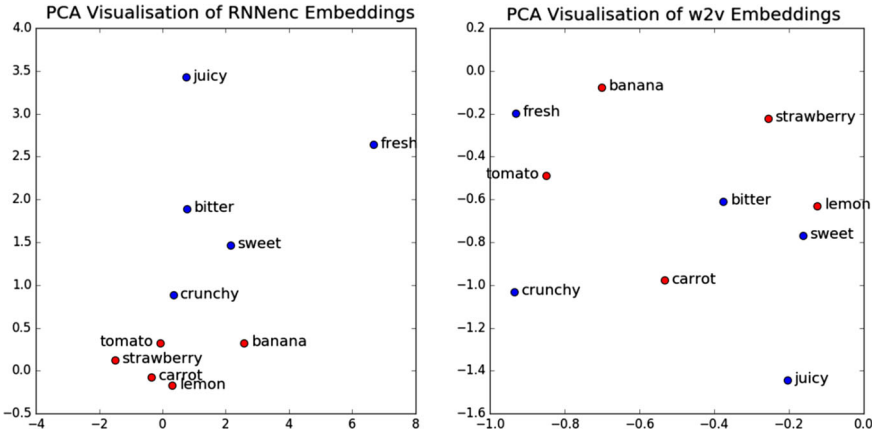

Fig. 1 A 2D PCA visualisation of word embedding spaces. Dimensions in the space correspond to the two principal components of the embedding space. The tendency of MT-based embeddings (RNNenc) to organise according to ontological similarity rather than topical associations can be observed on the *left*, where fruits and vegetables occupy one part of the space and associated qualities occupy another. In the Skipgram embedding space (*right*) this regularity is not observed

Table 2 Nearest neighbours (excluding plurals) in the embedding spaces of different models

	Skipgram	Glove	CW	FD	BiCVM	RNNenc	RNNsearch
Teacher	Vocational	Student	Student	Elementary	Faculty	Professor	Instructor
	In-service	Pupil	Tutor	School	Professors	Instructor	Professor
	College	University	Mentor	Classroom	Teach	Trainer	Educator
Eaten	Spoiled	Cooked	Baked	Ate	Eating	Ate	Ate
	Squeezed	Eat	Peeled	Meal	Eat	Consumed	Consumed
	Cooked	Eating	Cooked	Salads	Baking	Tasted	Eat
Britain	Northern	Ireland	Luxembourg	UK	UK	UK	England
	Great	Kingdom	Belgium	British	British	British	UK
	Ireland	Great	Madrid	London	England	America	Syria

All models were trained for six epochs on the translation corpus except CW and FD (as noted previously). NMT embedding spaces are oriented according to similarity, whereas embeddings learned by monolingual models are organized according to relatedness. The other bilingual model BiCVM also exhibits a notable focus on similarity

consequence of the fact, since it lacks an attention mechanism, a greater proportion of the memory capacity of the RNNenc model resides in its word embedding weights, which stimulates the acquisition of richer lexical representations.

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

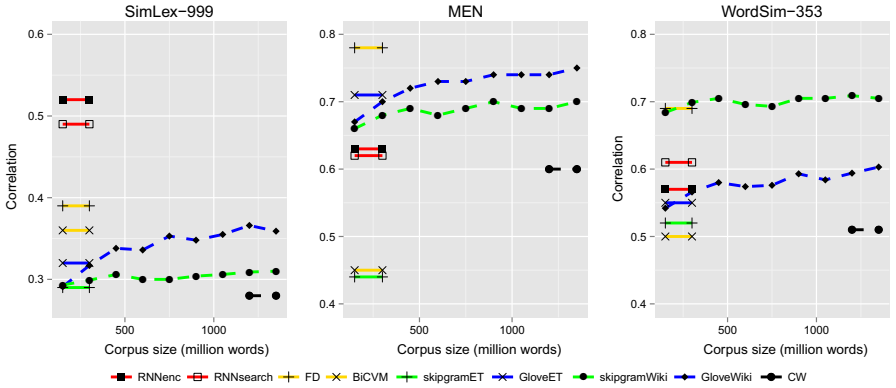

Fig. 2 The effect of increasing the amount of training data on the quality of monolingual embeddings, based on similarity-based evaluations (SimLex-999) and two relatedness-based evaluations (MEN and WordSim-353). *ET* in the legend indicates models trained on the English half of the translation corpus. *Wiki* indicates models trained on Wikipedia

to large quantities of training data was one of the principal motivations behind the skipgram architecture (Mikolov et al. 2013b). To check if monolingual models simply need more training data to capture similarity as effectively as bilingual models, we therefore trained them on increasingly large subsets of Wikipedia.[10] As shown in Fig. 2, this is not in fact the case. The performance of monolingual embeddings on similarity tasks remains well below the level of the NMT embeddings and somewhat lower than the non-MT bilingual embeddings as the amount of training data increases.

3.3 Analogy resolution

Lexical analogy questions have 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 surprisingly effective at answering such questions (Mikolov et al. 2013b). This is because, if $\mathbf{m}$, $\mathbf{b}$ and $\mathbf{w}$ are skipgram-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 embeddings on analogy questions using the same vector-algebra method as in (Mikolov et al. 2013b). As 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*),

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

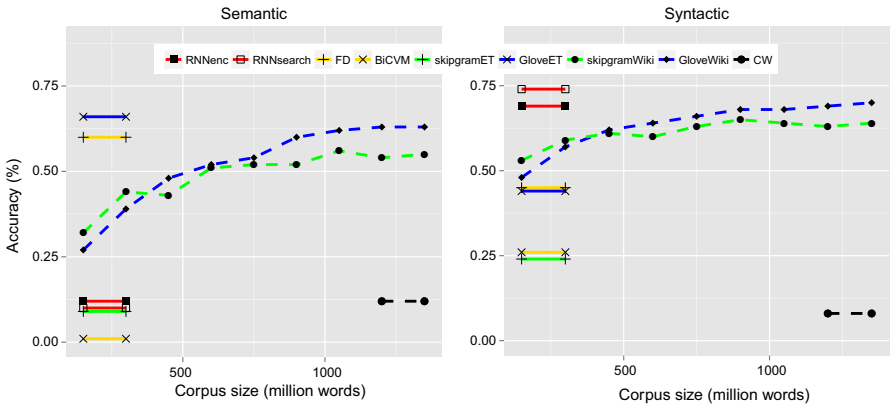

Fig. 3 Translation-based embeddings perform best on syntactic analogies (*run,ran:hide,hid*). Monolingual skipgram/Glove models are better at semantic analogies (*father, man; mother, woman*)

since successful answering seems to rely on some (world) knowledge of the concepts themselves.

As shown in Fig. 3, NMT embeddings yield relatively poor answers to semantic analogy questions compared with monolingual embeddings and the bilingual embeddings *FD* (which are projections of similar monolingual embeddings).[11] It appears that the translation objective prevents the embedding space from developing the same linear, geometric regularities as skipgram-style models with respect to semantic organisation. This also seems to be true of the embeddings from the full-sentence language model *CW*. Further, in the case of the Glove and FD models this advantage seems to be independent of both the domain and size of the training data, since embeddings from these models trained on only the English half of the translation corpus still outperform the translation embeddings.

On the other hand, NMT embeddings 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. The connection between the translation objective and the embedding of lexical syntactic information is further supported by the fact that embeddings learned by the bilingual model BiCVM do not perform comparably on the syntactic analogy task. In this model, sentential semantics is transferred via a bag-of-words representation, presumably rendering the precise syntactic information less important.

When considering the two properties of NMT embeddings highlighted by these experiments, namely the encoding of semantic similarity and lexical syntax, it is worth noting that items in the similarity-focused evaluations of the previous section (SimLex-

[11] The performance of the FD embeddings on this task is higher than that reported by Faruqui and Dyer (2014) because we search for answers over a smaller total candidate vocabulary.

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)

		EN–FR	EN–DE		'Earned'	'Castle'	'Money'
WordSim-353	ρ	0.60	**0.61**	**EN-FR**	*Gained*	*Chateau*	***Silver***
MEN	ρ	0.61	**0.62**		***Won***	*Palace*	*Funds*
SimLex-999	ρ	0.49	**0.50**		*Acquired*	*Fortress*	*Cash*
SimLex-Assoc-333	ρ	0.45	**0.47**				
TOEFL	%	0.90	**0.93**	**EN-DE**	*Gained*	*Chateau*	*Funds*
Syn/antonym	%	**0.72**	0.70		*Deserved*	*Palace*	*Cash*
Syntactic analogies	%	**0.73**	0.62		*Accummulated*	***Padlock***	*Resources*
Semantic analogies	%	0.10	**0.11**				

Bold italics indicate target-language-specific effects. Evaluation items and vocabulary searches were restricted to words common to both models

999 and TOEFL) consist of word groups or pairs that have identical syntactic role. Thus, even though lexical semantic information is in general pertinent to conceptual similarity (Levy and Goldberg 2014), the lexical syntactic and conceptual properties of translation embeddings are in some sense independent of one another.

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.

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 (348m 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 Overcoming the vocabulary size problem

A potential drawback to using NMT models for learning word embeddings is the computational cost of training such a model on large vocabularies. To generate a target language sentence, NMT models repeatedly compute a softmax distribution over the target vocabulary. This computation scales with vocabulary size and must be repeated for each word 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 NMT models as representation-learning tools. This contrasts with the shallower monolingual and bilingual representation-learning models considered in this paper, which efficiently compute a distribution over a large target vocabulary using either a hierarchical softmax (Morin and Bengio 2005) or approximate methods such as negative sampling (Mikolov et al. 2013b; Hermann and Blunsom 2014), and thus can learn large vocabularies of both source and target embeddings.

A recently proposed solution to this problem enables NMT 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. 2015). The algorithm uses (biased) importance sampling (Bengio and Sénécal 2003) to approximate the probability distribution of words over a large target vocabulary with a 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 NMT 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

Table 4 Comparison of embeddings learned by the original (RNN Search-30k English, French, 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 the intersection of all model vocabularies

		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**

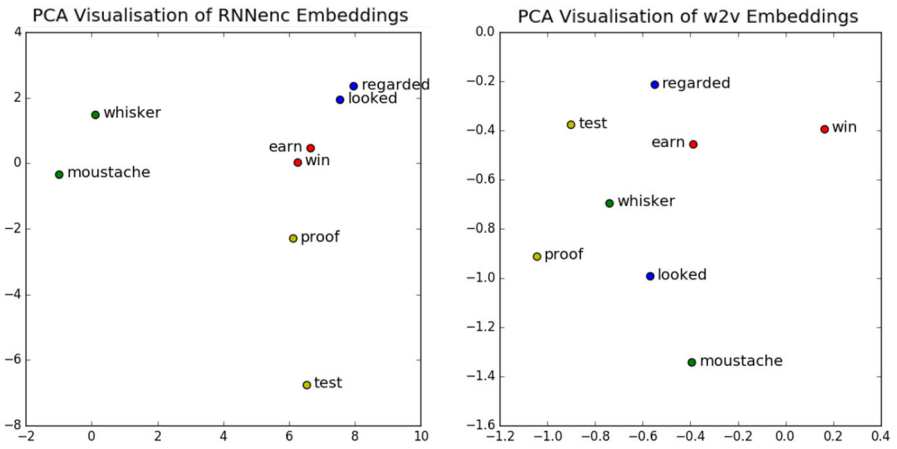

Fig. 4 The influence of the source-target language pair on the geometry of MT-based word embedding spaces. In the source embedding space of a model trained to translate from English to French (*left*), English word pairs that are generally translated to a single word in French (denoted by a *common colour*) occupy proximate locations. In the embedding space of Skipgram models trained on English text only (*right*), this effect is not observed

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/.[12]

6 How similarity emerges

Although NMT 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*. This effect is also observed in visualisations of the embedding spaces, as indicated in Fig. 4 .

These infrequent but striking differences between the English–German and English–French source embedding spaces indicate how similarity might emerge effectively in NMT 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

[12] A different solution to the rare-word problem was proposed by Luong et al. (2014). We do not evaluate the effects on the resulting embeddings of this method because we lack access to the source code.

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.

Such cases suggest that the induction of similarity in NMT models relies on the following condition on the semantic configuration between two languages.

(1) *For words* $\mathbf{w_1}$ *and* $\mathbf{w_2}$ *in the source language, there is some word* $\mathbf{t}$ *in the target language such that there are sentences in the training data in which* $\mathbf{w_1}$ *translates to* $\mathbf{t}$ *and sentences in which* $\mathbf{w_2}$ *translates to* $\mathbf{t}$.

if and only if

(2) $\mathbf{w_1}$ *and* $\mathbf{w_2}$ *are semantically similar.*

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* meaning either *win* or *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, $\mathbf{w_1}$ and $\mathbf{w_2}$, 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. This is also consistent with qualitative evidence presented by Faruqui and Dyer (2014) that projecting monolingual embeddings into a bilingual space orientates them to better reflect the synonymy/antonymy distinction.

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). Consistent with the qualitative observation made by Faruqui and Dyer (2014), we suggested how the nature of the semantic correspondence between the words in languages enables NMT embeddings to distinguish synonyms and antonyms and, more generally, to encode the information needed to reflect human intuitions of similarity.

The language-specific effects we observed in Sect. 4 suggest a potential avenue for improving translation and multi-lingual embeddings in future work. First, as hardware improves, training speeds fall, and data becomes more prevalent, we would like to explore the embeddings learned by NMT models that translate between much more distant language pairs such as English–Chinese or English–Arabic. For these language

pairs, the word alignment will be less monotonic and may result in even more important semantic and syntactic information being encoded in the lexical representations. Further, as observed by both Hermann and Blunsom (2014) and Faruqui and Dyer (2014), the bilingual representation learning paradigm can be naturally extended to update representations based on correspondences between multiple languages (for instance by interleaving English–French and English–German training examples). Such an approach should smooth out language-specific effects, leaving embeddings that encode only language-agnostic conceptual semantics and are thus more generally applicable. Another related challenge is to develop smaller or less complex representation-learning tools that encode similarity with as much fidelity as NMT models but without either the computational overhead or the requirement for sentence-aligned parallel corpora. Such a development would enable the techniques proposed here to be used to learn word representations for a wider range of low-resource languages.

Not all word embeddings learned from text are born equal. Depending on the application, those learned by NMT models may have particularly desirable properties. 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 humah-quality concept representations. 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 for extracting word meanings, is only sometimes true.

Acknowledgements This work was in part funded by a Google European Doctoral Fellowship and a Google Faculty Award.

References

Agirre E, Alfonseca E, Hall K, Kravalova J, Pasca M, Soroa A (2009) A study on similarity and relatedness using distributional and wordnet-based approaches. In: Proceedings of NAACL-HLT 2009

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of ICLR

Baroni M, Dinu G, Kruszewski G (2014) Dont 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, vol 1

Bengio Y, Sénécal JS (2003) Quick training of probabilistic neural nets by importance sampling. In: Proceedings of AISTATS 2003

Bengio Y, Ducharme R, Vincent P, Janvin C (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155

Bruni E, Tran NK, Baroni M (2014) Multimodal distributional semantics. J Artif Intell Res(JAIR) 49:1–47

Chandar S, Lauly S, Larochelle H, Khapra MM, Ravindran B, Raykar V, Saha A (2014) An autoencoder approach to learning bilingual word representations. In: NIPS

Cho K, van Merrienboer B, Gulcehre C, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the empirical methods in natural language processing (EMNLP 2014), to appear

Collobert R, Weston J (2008) A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th international conference on machine learning, ACM, pp 160–167

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. J Mach Learn Res 12:2493–2537

Devlin J, Zbib R, Huang Z, Lamar T, Schwartz R, Makhoul J (2014) Fast and robust neural network joint models for statistical machine translation. In: 52nd annual meeting of the association for computational linguistics, Baltimore, June

Faruqui M, Dyer C (2014) Improving vector space word representations using multilingual correlation. In: Proceedings of EACL, vol 2014

Firth RJ (1957) A synopsis of linguistic theory 1930–1955. Philological Society, Oxford, pp 1–32

Haghighi A, Liang P, Berg-Kirkpatrick T, Klein D (2008) Learning bilingual lexicons from monolingual corpora. In: ACL, vol 2008, pp 771–779

Hermann KM, Blunsom P (2014) Multilingual distributed representations without word alignment. In: Proceedings of ICLR

Hill F, Korhonen A (2014) Learning abstract concepts from multi-modal data: since you probably can't see what i mean. In: Proceedings of the empirical methods in natural language processing (EMNLP 2014)

Hill F, Reichart R, Korhonen A (2014) Simlex-999: evaluating semantic models with (genuine) similarity estimation. arXiv preprint arXiv:1408.3456

Jean S, Cho K, Memisevic R, Bengio Y (2015) On using very large target vocabulary for neural machine translation. In: Proceedings of NAACL

Kalchbrenner N, Blunsom P (2013) Recurrent continuous translation models. In: Proceedings of the 2013 conference on empirical methods in natural language processing, Association for Computational Linguistics, Seattle

Klementiev A, Titov I, Bhattarai B (2012a) Inducing crosslingual distributed representations of words. COLING

Klementiev A, Titov I, Bhattarai B (2012b) Inducing crosslingual distributed representations of words. In: COLING

Kočiský T, Hermann KM, Blunsom P (2014) Learning bilingual word representations by marginalizing alignments. In: Proceedings of ACL

Kusner M, Sun Y, Kolkin N, Weinberger KQ (2015) From word embeddings to document distances. In: Proceedings of the 32nd international conference on machine learning (ICML-15), pp 957–966

Landauer TK, Dumais ST (1997) A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev 104(2):211

Levy O, Goldberg Y (2014) Dependency-based word embeddings. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, vol 2

Luong T, Sutskever I, Le QV, Vinyals O, Zaremba W (2014) Addressing the rare word problem in neural machine translation. arXiv preprint arXiv:1410.8206

Mikolov T, Le QV, Sutskever I (2013a) Exploiting similarities among languages for machine translation. In: CORR

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013b) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119

Mnih A, Hinton GE (2009) A scalable hierarchical distributed language model. In: Advances in neural information processing systems, pp 1081–1088

Morin F, Bengio Y (2005) Hierarchical probabilistic neural network language model. AISTATS, Citeseer 5:246–252

Nelson DL, McEvoy CL, Schreiber TA (2004) The university of south florida free association, rhyme, and word fragment norms. Behav Res Methods Instrum Comput 36(3):402–407

Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: Proceedings of the empirical methods in natural language processing (EMNLP 2014)

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Proceedings of NIPS

Turney PD, Pantel P (2010) From frequency to meaning: vector space models of semantics. J Artif Intell Res 37(1):141–188

Vulić I, De Smet W, Moens MF (2011) Identifying word translations from comparable corpora using latent topic models. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers, Vol 2, Association for Computational Linguistics, pp 479–484

Weston J, Bengio S, Usunier N (2010) Large scale image annotation: learning to rank with joint word-image embeddings. Mach Learn 81(1):21–35