Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

EMNLP 2014

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau,
Fethi Bougares, Holger Schwenk, Yoshua Bengio

Universit´e de Montr´eal

Universite du Maine

Jacobs University

Part 1.

Background

- **The field of statistical machine translation (SMT)**
  - Deep neural networks have begun to show promising results
    - These are not limited to language modeling, paraphrase detection, word embedding extraction
  - (Schwenk, 2012)
    - Summarizes a successful usage of feed forward neural networks in the framework of phrase-based SMT system
  - Problem Definition
    - Novel neural network architecture that can be used as a part of the conventional phrase-based SMT system

Introduction

- **RNN Encoder–Decoder**
  - Consists of recurrent neural networks (RNN) that act as an encoder and a decoder pair
  - The networks are trained jointly to maximize the conditional probability of the target sequence given a source sequence

  - Encoder
    - Map a variable-length source sequence to a fixed-length vector

  - Decoder
    - Map the vector representation back to a variable-length target sequence

  - Use a rather sophisticated hidden unit
    - Improve both the memory capacity and the ease of training

Introduction

- **RNN Encoder–Decoder**
  - Improve the translation performance (English to French)
    - The model is then used as a part of a standard phrase-based SMT system by scoring each phrase pair in the phrase table

  - Better at capturing the linguistic regularities in the phrase table
    - Analyze the trained RNN Encoder–Decoder by comparing its phrase scores with those given by the existing translation model
    - Explain the quantitative improvements in the overall translation performance

  - Learn a continuous space representation of a phrase that preserves both the semantic and syntactic structure of the phrase

Preliminary: Recurrent Neural Networks

- **Recurrent neural network (RNN)**
  - Consist of a hidden state h and an optional output y which operates on a variable length sequence $\mathbf{x} = (x_1, \ldots, x_T)$
  - At each time step t, the hidden state $\mathbf{h}_{\langle t \rangle}$

    RNN is updated by $\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, x_t\right)$

    A non-linear activation function f

Preliminary: Recurrent Neural Networks

- **Recurrent neural network (RNN)**
  - Learn a probability distribution over a sequence by being trained to predict the next symbol in a sequence

Conditional distribution

$$p(x_{t,j} = 1 \mid x_{t-1}, \ldots, x_1) = \frac{\exp\left(\mathbf{w}_j \mathbf{h}_{\langle t \rangle}\right)}{\sum_{j'=1}^{K} \exp\left(\mathbf{w}_{j'} \mathbf{h}_{\langle t \rangle}\right)}$$

All possible symbols

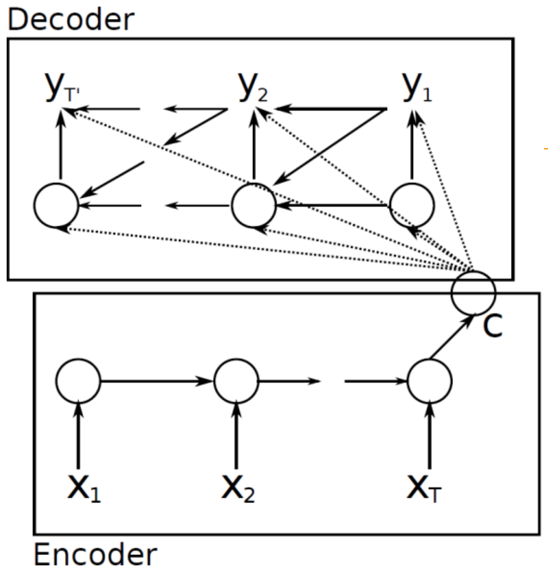
$$j = 1, \ldots, K$$

  - By combining these probabilities, we can compute the probability of the sequence x

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_{t-1}, \ldots, x_1)$$

  - From this learned distribution, Sample a new sequence by iteratively sampling a symbol at each time step

RNN Encoder-Decoder



- **Recurrent neural network (RNN)**
  - Encoder
    - A variable-length sequence into a fixed-length vector representation
    - Reads each symbol of an input sequence x sequentially
  - Decoder
    - A Fixed-length vector representation back into a variable-length sequence
    - Generate the output sequence by predicting the next symbol y_t given the hidden state $\mathbf{h}_{\langle t \rangle}$
    - y_t $\mathbf{h}_{\langle t \rangle}$ are also conditioned on y_{t-1} and on the summary $\mathbf{c}$ of the input sequence
    - The hidden state of the decoder

$$\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, y_{t-1}, \mathbf{c}\right)$$

    - The conditional distribution of the next symbol

$$P(y_t | y_{t-1}, y_{t-2}, \ldots, y_1, \mathbf{c}) = g\left(\mathbf{h}_{\langle t \rangle}, y_{t-1}, \mathbf{c}\right)$$

    - Encoder–decoder are jointly trained to maximize the conditional log-likelihood

$$\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{y}_n \mid \mathbf{x}_n)$$

RNN Encoder-Decoder

- **Training**
  - Generate a target sequence given an input sequence
  - Score a given pair of input and output sequences where the score is simply a probability $p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$

Hidden Unit that Adaptively Remembers and Forgets

- **Describe how the activation of the j-th hidden unit is computed**
  - **Reset gate** $$r_j = \sigma\left(\left[\mathbf{W}_r\mathbf{x}\right]_j + \left[\mathbf{U}_r\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

    Logistic sigmoid function σ

    The j-th element of a vector $[.]_j$
  - **Update gate** $$z_j = \sigma\left(\left[\mathbf{W}_z\mathbf{x}\right]_j + \left[\mathbf{U}_z\mathbf{h}_{\langle t-1\rangle}\right]_j\right)$$

    Weight matrices which are learned $\mathbf{W}_r$ $\mathbf{U}_r$
  - **The actual activation** $$h_j^{\langle t\rangle} = z_j h_j^{\langle t-1\rangle} + (1 - z_j)\tilde{h}_j^{\langle t\rangle}$$

    Weight matrices which are learned $$\tilde{h}_j^{\langle t\rangle} = \phi\left(\left[\mathbf{W}\mathbf{x}\right]_j + \left[\mathbf{U}\left(\mathbf{r}\odot\mathbf{h}_{\langle t-1\rangle}\right)\right]_j\right)$$

Hidden Unit that Adaptively Remembers and Forgets

- **Training**
  - When the reset gate is close to 0
    - The hidden state is forced to ignore the previous hidden state and reset with the current input only
    - Effectively allow the hidden state to drop any information that is found to be irrelevant later in the future
    - Allow a more compact representation
  - The update gate controls how much information from the previous hidden state will carry over to the current hidden state
    - Act similarly to the memory cell in the LSTM network
    - Help the RNN to remember long term information
    - This may be considered an adaptive variant of a leaky-integration unit (Bengio et al., 2013)

$$h_{t,i} = \alpha_i h_{t-1,i} + (1 - \alpha_i) F_i(h_{t-1}, x_t)$$

$$n_{leaky} \in \{0\%, 25\%, 50\%\}$$

Luyu Gao et al. Unsupervised corpus aware language model pre-training for dense passage retrieval. ACL, 2022.

Hidden Unit that Adaptively Remembers and Forgets

- **RNN Encoder**
  - Source phrase $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ / Target phrase $Y = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M)$
    - Source phrase is embedded in a 500-dimensional vector space $e(\mathbf{x}_i) \in \mathbb{R}^{500}$
  - The hidden state

$$h_j^{\langle t \rangle} = z_j h_j^{\langle t-1 \rangle} + (1 - z_j)\tilde{h}_j^{\langle t \rangle}$$

$$\tilde{h}_j^{\langle t \rangle} = \tanh\left(\left[\mathbf{W}e(\mathbf{x}_t)\right]_j + \left[\mathbf{U}\left(\mathbf{r} \odot \mathbf{h}_{\langle t-1 \rangle}\right)\right]_j\right)$$

$$z_j = \sigma\left(\left[\mathbf{W}_z e(\mathbf{x}_t)\right]_j + \left[\mathbf{U}_z \mathbf{h}_{\langle t-1 \rangle}\right]_j\right)$$

$$r_j = \sigma\left(\left[\mathbf{W}_r e(\mathbf{x}_t)\right]_j + \left[\mathbf{U}_r \mathbf{h}_{\langle t-1 \rangle}\right]_j\right)$$

  - σ Logistic sigmoid function / $\odot$ Element-wise multiplication
  - The representation of the source phrase $\mathbf{c} = \tanh\left(\mathbf{V}\mathbf{h}^{\langle N \rangle}\right)$

Hidden Unit that Adaptively Remembers and Forgets

- **RNN Decoder**
  - The initial hidden state $\mathbf{h}'^{\langle 0 \rangle} = \tanh\left(\mathbf{V}'\mathbf{c}\right)$
    - Use $'$ to distinguish parameters of the decoder from those of the encoder
  - The hidden state

$$h'^{\langle t \rangle}_j = z'_j h'^{\langle t-1 \rangle}_j + (1 - z'_j)\tilde{h}'^{\langle t \rangle}_j$$

$$\tilde{h}'^{\langle t \rangle}_j = \tanh\left(\left[\mathbf{W}' e(\mathbf{y}_{t-1})\right]_j + r'_j \left[\mathbf{U}'\mathbf{h}'_{\langle t-1 \rangle} + \mathbf{C}\mathbf{c}\right]\right)$$

$$r'_j = \sigma\left(\left[\mathbf{W}'_r e(\mathbf{y}_{t-1})\right]_j + \left[\mathbf{U}'_r \mathbf{h}'_{\langle t-1 \rangle}\right]_j + \left[\mathbf{C}_r \mathbf{c}\right]_j\right)$$

  - $e(\mathbf{y}_0)$ An all-zero vector
  - $e(\mathbf{y})$ An embedding of a target word (similarly to the case of the encoder)
  - i-element of $\mathbf{s}_{\langle t \rangle}$ is $s_i^{\langle t \rangle} = \max\left\{s'^{\langle t \rangle}_{2i-1}, s'^{\langle t \rangle}_{2i}\right\}$

$$\mathbf{s}'^{\langle t \rangle} = \mathbf{O}_h \mathbf{h}'^{\langle t \rangle} + \mathbf{O}_y \mathbf{y}_{t-1} + \mathbf{O}_c \mathbf{c}$$

Hidden Unit that Adaptively Remembers and Forgets

- **RNN Decoder**
  - The probability of generating j-th word

$$p(y_{t,j} = 1 \mid \mathbf{y}_{t-1}, \ldots, \mathbf{y}_1, X) = \frac{\exp\left(\mathbf{g}_j \mathbf{s}_{\langle t \rangle}\right)}{\sum_{j'=1}^{K} \exp\left(\mathbf{g}_{j'} \mathbf{s}_{\langle t \rangle}\right)}$$

  - i-element of $\mathbf{s}_{\langle t \rangle}$

$$s_i^{\langle t \rangle} = \max\left\{ {s'}_{2i-1}^{\langle t \rangle}, {s'}_{2i}^{\langle t \rangle} \right\}$$

$$\mathbf{s'}^{\langle t \rangle} = \mathbf{O}_h \mathbf{h'}^{\langle t \rangle} + \mathbf{O}_y \mathbf{y}_{t-1} + \mathbf{O}_c \mathbf{c}$$

  - A so-called maxout unit $s_i^{\langle t \rangle}$
  - Computational efficiency: instead of a single-matrix output weight G
  - Use a product of two matrices such that $\mathbf{G} = \mathbf{G}_l \mathbf{G}_r$

$\mathbf{G}_l \in \mathbb{R}^{K \times 500}$ $\mathbf{G}_r \in \mathbb{R}^{500 \times 1000}$

Statistical Machine Translation

- **The goal of the system**
  - Find a translation f given a source sentence e
    - Maximize $p(\mathbf{f} \mid \mathbf{e}) \propto p(\mathbf{e} \mid \mathbf{f}) p(\mathbf{f})$
    - RIght hand side is called translation model and the latter language model
  - Log linear model with features & weights $\log p(\mathbf{f} \mid \mathbf{e})$

$$\log p(\mathbf{f} \mid \mathbf{e}) = \sum_{n=1}^{N} w_n f_n(\mathbf{f}, \mathbf{e}) + \log Z(\mathbf{e})$$

    - f_n and w_n are the n-th feature and weight, respectively
    - Normalization constant $Z(\mathbf{e})$

Statistical Machine Translation

- **Phrase-based SMT framework**
  - Translation model $\log p(\mathbf{e} \mid \mathbf{f})$
    - It is factorized into the translation probabilities of matching phrases in the source and target sentences
  - Neural networks have been used widely in SMT systems
    - Neural networks have been used to rescore translation hypotheses (n-best lists)
    - Score the translated sentence (or phrase pairs) using a representation of the source sentence as an additional input

Statistical Machine Translation

- **Scoring Phrase Pairs with RNN Encoder–Decoder**
  - Training:
  - Ignore the (normalized) frequencies of each phrase pair in the original corpora
    - Reduce the computational expense of randomly selecting phrase pairs from a large phrase table according to the normalized frequencies
    - Ensure that the RNN Encoder-Decoder does not simply learn to rank the phrase pairs according to their numbers of occurrences
  - Existing translation probability in the phrase table
    - Reflects the frequencies of the phrase pairs in the original corpus
  - With a fixed capacity of the RNN Encoder–Decoder
    - Try to ensure that most of the capacity of the model is focused toward learning linguistic regularities
    - i.e., distinguishing between plausible and implausible translations, or learning the "manifold" (region of probability concentration) of plausible translations

- **Scoring Phrase Pairs with RNN Encoder–Decoder**
  - After Training:
  - Add a new score for each phrase pair to the existing phrase table
    - Enter into the existing tuning algorithm with minimal additional overhead in computation
  - Possible to completely replace the existing phrase table (Schwenk, 2012)
    - For a given source phrase, the RNN Encoder–Decoder will need to generate a list of (good) target phrases
    - Requires an expensive sampling procedure to be performed repeatedly
    - Only consider rescoring the phrase pairs in the phrase table

Holger Schwenk. Continuous space translation models for phrase-based statistical machine translation. COLING. 2012

Experiment

- **Experimental setting**
  - Task: English/French translation on WMT'14 workshop
  - All the out-of-vocabulary words were mapped to a special token ([UNK])
  - Data source
    - The bilingual corpora include Europarl (61M words), news commentary (5.5M), UN (421M)
    - Two crawled corpora of 90M and 780M words
  - The most relevant subset of the data for a given task
    - Language modeling: A subset of 418M words out of more than 2G words
    - The RNN Encoder–Decoder: A subset of 348M out of 850M words
  - Test set
    - newstest2012 and 2013 for data selection, newstest2014
    - weight tuning with MERT
  - Training set
    - Limit the source & target vocabulary to the most frequent 15,000 words
    - Cover approximately 93% of the dataset

Experiment

- **RNN Encoder–Decoder**
  - 1000 hidden units with the proposed gates at the encoder and at the decoder
  - The input matrix between each input symbol $x_{\langle t \rangle}$ and the hidden unit is approximated with two lower-rank matrices
  - The output matrix is approximated similarly
  - Use rank-100 matrices, equivalent to learning an embedding of dimension 100 for each word
  - From the hidden state in the decoder to the output:

    A deep neural network (Pascanu et al., 2014) with a single intermediate layer having 500 maxout units each pooling 2 inputs

Experiment

- **Neural Language Model**
  - All the weight parameters were initialized by sampling from an isotropic zero-mean (white) Gaussian distribution with its standard deviation fixed to 0:01, except for the recurrent weight parameters
  - For the recurrent weight matrices, we first sampled from a white Gaussian distribution and used its left singular vectors matrix, following (Saxe et al., 2014)
  - At each update, we used 64 randomly selected phrase pairs from a phrase table (which was created from 348M words)

Experiment

- **Quantitative Analysis**
  - Contributions of the CSLM and the RNN Encoder-Decoder are not too correlated and that one can expect better results by improving each method independently
  - Penalize the number of words that are unknown to the neural networks (i.e. words which are not in the shortlist)

Models	BLEU	
	dev	test
Baseline	30.64	33.30
RNN	31.20	33.87
CSLM + RNN	31.48	34.64
CSLM + RNN + WP	31.50	34.54

WP (a word penalty):
Penalize the number of unknown words to neural networks

Conditional probability of any $x_t^i \notin$ SL

$$p\left(x_t = [\mathrm{UNK}] \mid x_{<t}\right) = p\left(x_t \notin \mathrm{SL} \mid x_{<t}\right) = \sum_{x_t^j \notin SL} p\left(x_t^j \mid x_{<t}\right) \geq p\left(x_t^i \mid x_{<t}\right)$$

- $x_{<t}$ is a shorthand notation for $x_{t-1}, \ldots, x_1$

Experiment

- **Qualitative Analysis**
  - Analyze the phrase pair scores
  - The existing translation model relies solely on the statistics of the phrase pairs in the corpus
    - Expect its scores to be better estimated for the frequent phrases but badly estimated for rare phrases
    - Further expect model which was trained without any frequency information to score the phrase pairs based rather on the linguistic regularities
  - Focus on those pairs whose source phrase is long (more than 3 words per source phrase) and frequent
  - For each such source phrase,

  look at the target phrases that have been scored high either by the translation probability $p(\mathbf{f} \mid \mathbf{e})$ or by the RNN Encoder–Decoder

Experiment

• Qualitative Analysis

Top-3 target phrases per source phrase

- Source phrases were randomly selected from phrases with 4 or more words
- **?** denotes an incomplete (partial) character
- **г** is a Cyrillic letter ghe

Source	Translation Model	RNN Encoder–Decoder
at the end of the	[a la fin de la] [ŕ la fin des années] [être supprimés à la fin de la]	[à la fin du] [à la fin des] [à la fin de la]
for the first time	[**г** © pour la premirëre fois] [été donnés pour la première fois] [été commémorée pour la première fois]	[pour la première fois] [pour la première fois ,] [pour la première fois que]
in the United States and	[? aux ?tats-Unis et] [été ouvertes aux États-Unis et] [été constatées aux États-Unis et]	[aux Etats-Unis et] [des Etats-Unis et] [des États-Unis et]
, as well as	[?s , qu'] [?s , ainsi que] [?re aussi bien que]	[, ainsi qu'] [, ainsi que] [, ainsi que les]
one of the most	[?t ?l' un des plus] [?l' un des plus] [être retenue comme un de ses plus]	[l' un des] [le] [un des]

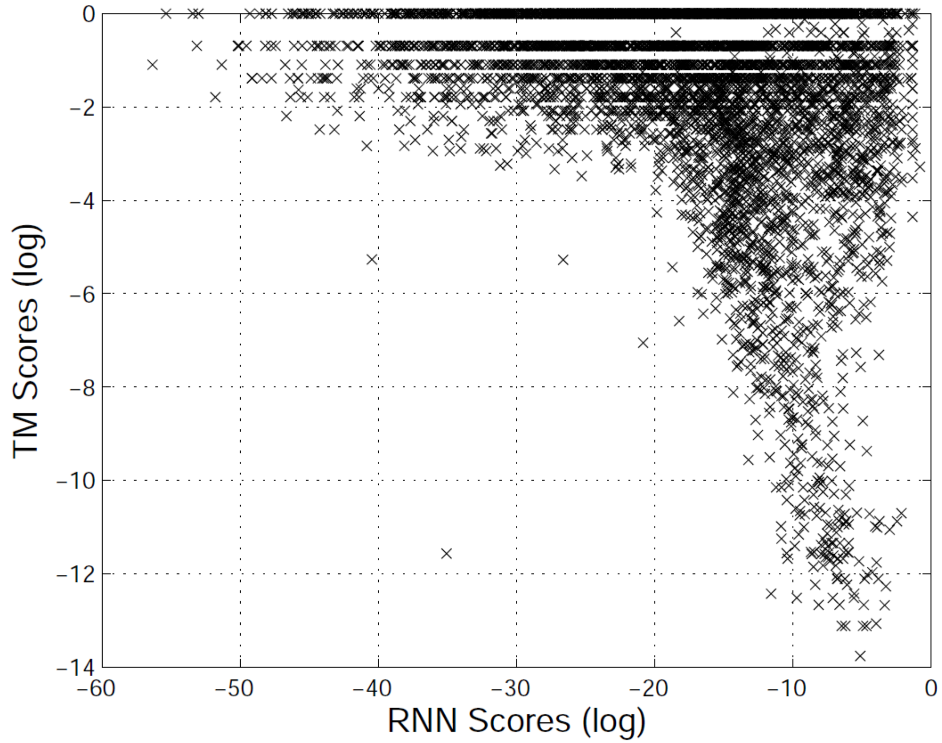
(a) Long, frequent source phrases

Source	Translation Model	RNN Encoder–Decoder
, Minister of Communications and Transport	[Secrétaire aux communications et aux transports :] [Secrétaire aux communications et aux transports]	[Secrétaire aux communications et aux transports] [Secrétaire aux communications et aux transports :]
did not comply with the	[vestimentaire , ne correspondaient pas à des] [susmentionnée n' était pas conforme aux] [présentées n' étaient pas conformes à la]	[n' ont pas respecté les] [n' était pas conforme aux] [n' ont pas respecté la]
parts of the world .	[© gions du monde .] [régions du monde considérées .] [région du monde considérée .]	[parties du monde .] [les parties du monde .] [des parties du monde .]
the past few days .	[le petit texte .] [cours des tout derniers jours .] [les tout derniers jours .]	[ces derniers jours .] [les derniers jours .] [cours des derniers jours .]
on Friday and Saturday	[vendredi et samedi à la] [vendredi et samedi à] [se déroulera vendredi et samedi ,]	[le vendredi et le samedi] [le vendredi et samedi] [vendredi et samedi]

(b) Long, rare source phrases

Experiment

- **Qualitative Analysis**
  - Many other phrase pairs that were scored radically different
  - This could arise from the proposed approach of training the RNN
  - Discourage the RNN from learning simply the frequencies of the phrase pairs from the corpus, as explained earlier



The visualization of phrase pairs according to their scores

- Log-probabilities
- The RNN Encoder–Decoder and the translation model

Experiment

• Qualitative Analysis

- The generated phrases do not overlap completely with the target phrases from the phrase table
- Encourage us to further investigate the possibility of replacing the whole or a part of the phrase table with the proposed RNN Encoder–Decoder in the future
- The top-5 target phrases out of 50 samples sorted by the RNN scores

Samples generated from the RNN Encoder–Decoder

- The top-5 target phrases out of 50 samples
- They are sorted by the RNN Encoder-Decoder scores

Source	Samples from RNN Encoder–Decoder
at the end of the	[à la fin de la] (×11)
for the first time	[pour la première fois] (×24) [pour la première fois que] (×2)
in the United States and	[aux États-Unis et] (×6) [dans les États-Unis et] (×4)
, as well as	[, ainsi que] [,] [ainsi que] [, ainsi qu'] [et UNK]
one of the most	[l' un des plus] (×9) [l' un des] (×5) [l' une des plus] (×2)

(a) Long, frequent source phrases

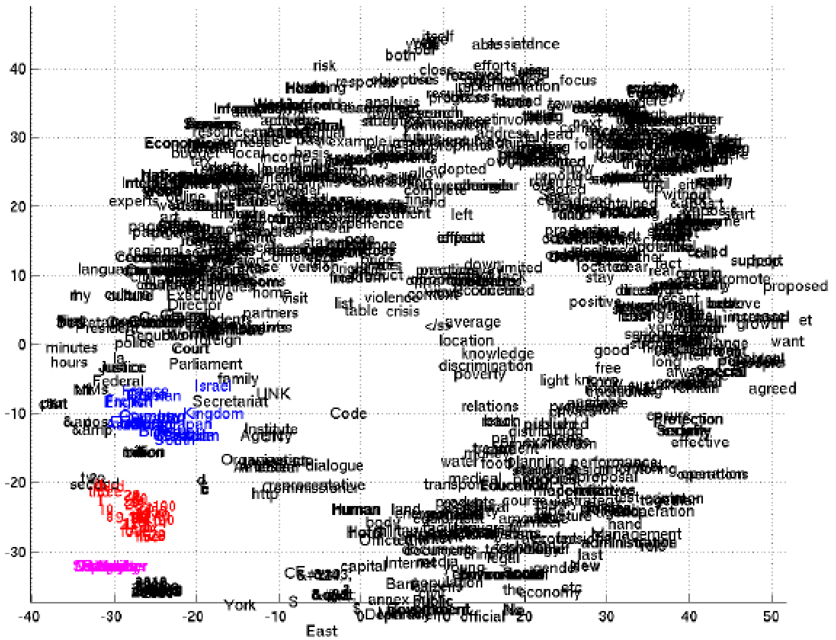
Source	Samples from RNN Encoder–Decoder
, Minister of Communications and Transport	[, ministre des communications et le transport] (×13)
did not comply with the	[n' tait pas conforme aux] [n' a pas respect l'] (×2) [n' a pas respect la] (×3)
parts of the world .	[arts du monde .] (×11) [des arts du monde .] (×7)
the past few days .	[quelques jours .] (×5) [les derniers jours .] (×5) [ces derniers jours .] (×2)
on Friday and Saturday	[vendredi et samedi] (×5) [le vendredi et samedi] (×7) [le vendredi et le samedi] (×4)

(b) Long, rare source phrases

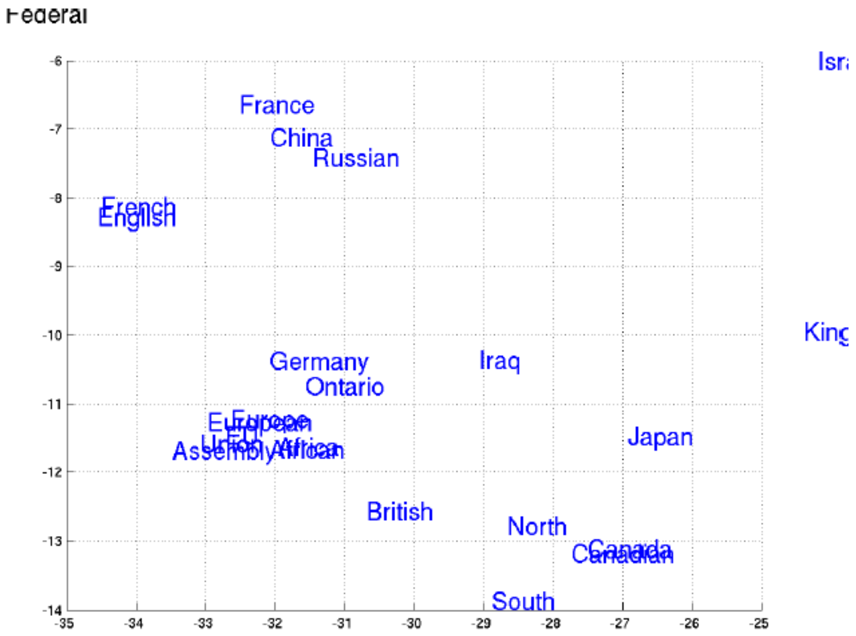
Experiment

- **Qualitative Analysis**
  - Continuous space language models using neural networks are able to learn semantically meaningful embeddings
  - Visualize the representations of the phrases that consists of four or more words using the Barnes-Hut-SNE
  - Encoder-Decoder captures both semantic and syntactic structures of the phrases

The full embedding space



A zoomed-in view of one region (color–coded)



Conclusion

- **Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation**
  - A new neural network architecture, called an RNN Encoder–Decoder
    - A novel hidden unit that includes a reset gate and an update gate that adaptively control how much each hidden unit remembers or forgets while reading/generating a sequence
  - Able to learn the mapping from a sequence of an arbitrary length to another sequence, possibly from a different set, of an arbitrary length
  - Able to either score a pair of sequences (in terms of a conditional probability) or generate a target sequence given a source sequence
  - Evaluate the proposed model with the task of statistical machine translation
    - The new model is able to capture linguistic regularities in the phrase pairs well
    - Improve the overall translation performance in terms of BLEU scores
    - Rather orthogonal to the existing approach of using neural networks in the SMT system
  - Captures the linguistic regularities in multiple levels
    - i.e. at the word level as well as phrase level

Conclusion

- **Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation**
    - One approach that was not investigated here is to replace the whole, or a part of the phrase table by letting the RNN Encoder–Decoder propose target phrases
    - Noting that the proposed model is not limited to being used with written language, it will be an important future research to apply the proposed architecture to other applications such as speech transcription

Appendix

- **Spanning Longer Time Ranges with Leaky Integration unit (Bengio et al., 2013)**
  - Long-Short-Term Memory (LSTM) networks handling much longer range dependencies
    - Benefit from a linearly self-connected memory unit with a near 1 self-weight
    - A near 1 self-weight allows signals (and gradients) to propagate over long time spans
  - A different interpretation to this slow-changing units is that they behave like low-pass filter
  - Hence they can be used to focus certain units on different frequency regions of the data
  - Band-pass filter units

    : Passes frequencies within a certain range and rejects frequencies outside that range
    - Decide on what frequency bands different units should focus
    - Add low frequency information as an additional input to a recurrent network helps improving the performance of the model

Luyu Gao et al. Unsupervised corpus aware language model pre-training for dense passage retrieval. ACL, 2022.

Appendix

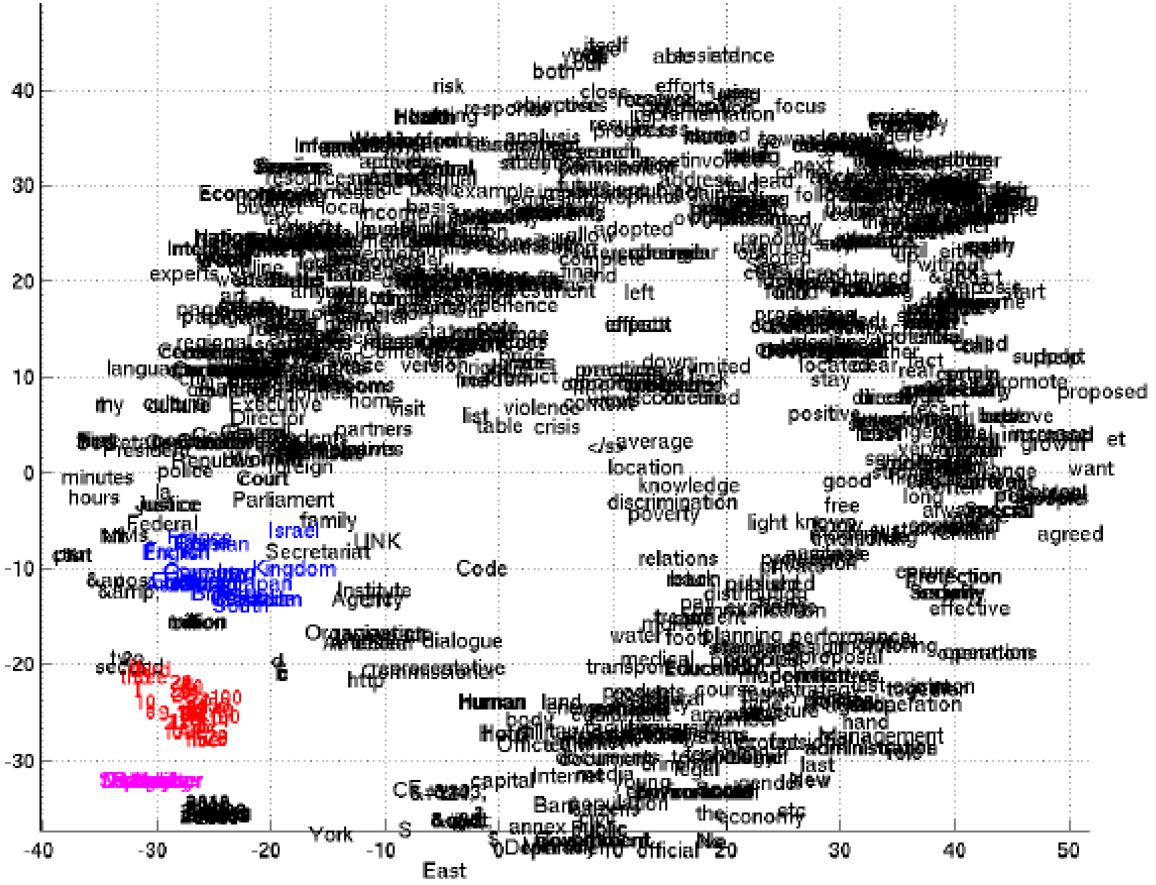
- **Leaky-integration unit (Bengio et al., 2013)**
  - State-to-state map $h_{t,i} = \alpha_i h_{t-1,i} + (1 - \alpha_i) F_i(h_{t-1}, x_t)$
  - The standard RNN corresponds to $\alpha_i = 0$
  - Different values of αi were randomly sampled from (0.02, 0.2)
  - Allow some units to react quickly while others are forced to change slowly
  - But also propagate signals and gradients further in time
  - Leaky factors $\alpha < 1$
    - The vanishing effect is still present
    - But the time-scale of the vanishing effect can be expanded

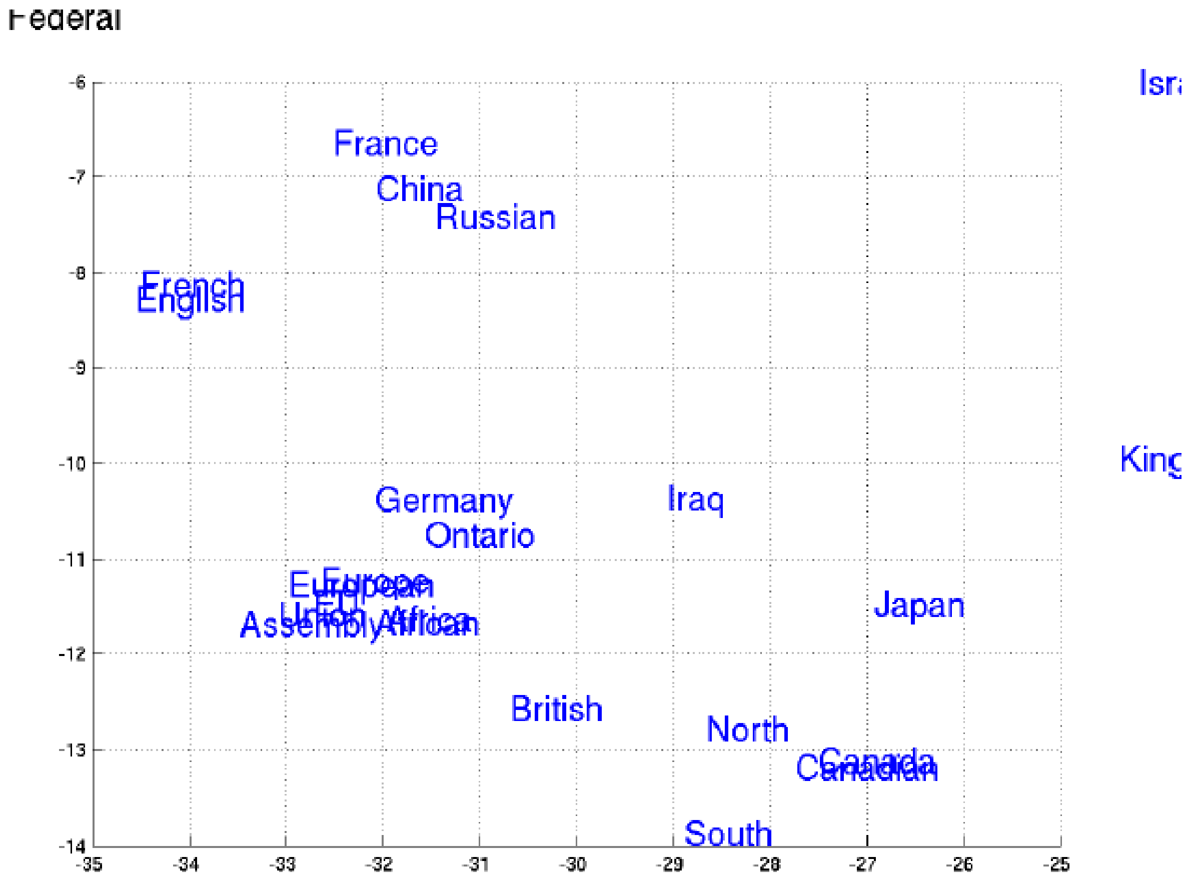
$$n_{leaky} \in \{0\%, 25\%, 50\%\}$$

Luyu Gao et al. Unsupervised corpus aware language model pre-training for dense passage retrieval. ACL, 2022.

Appendix

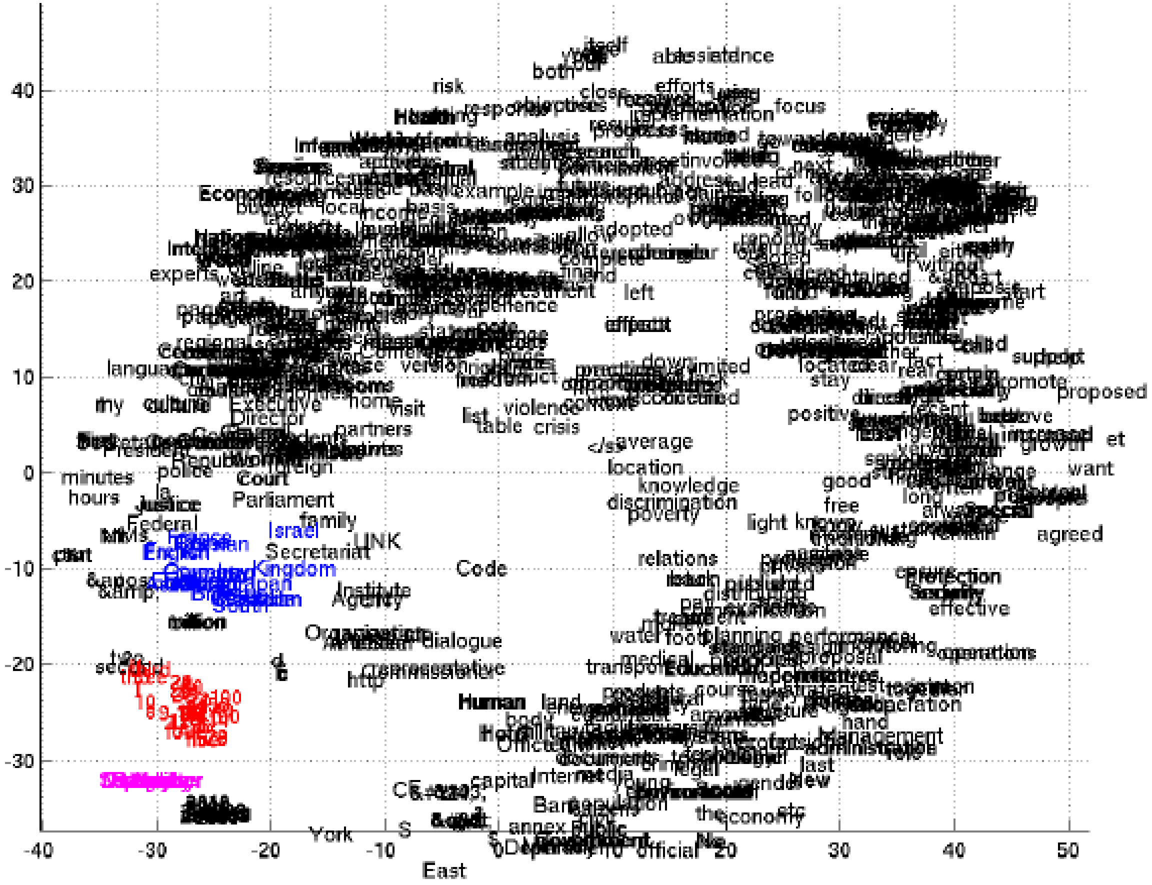
- **2–D embedding of the learned word representation**
    - The left one shows the full embedding space, while the other figures show the zoomed-in view of specific regions (color–coded)

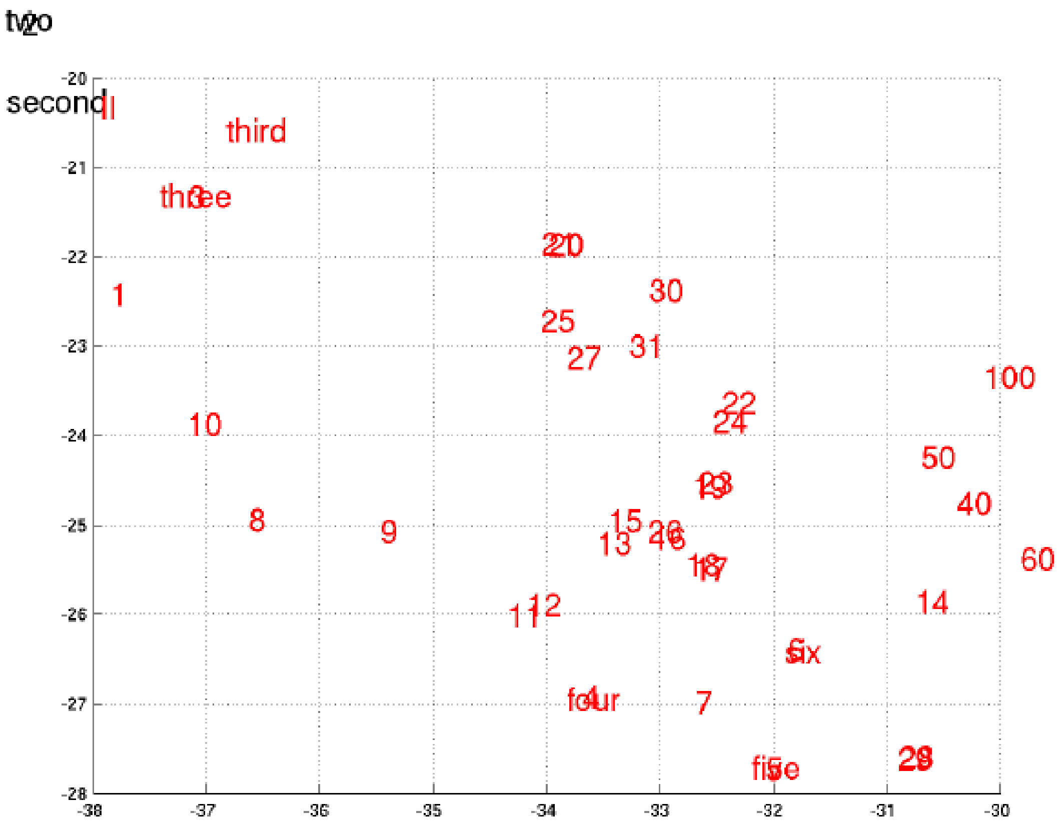




Appendix

- **2–D embedding of the learned word representation**
  - The left one shows the full embedding space, while the other figures show the zoomed-in view of specific regions (color–coded)

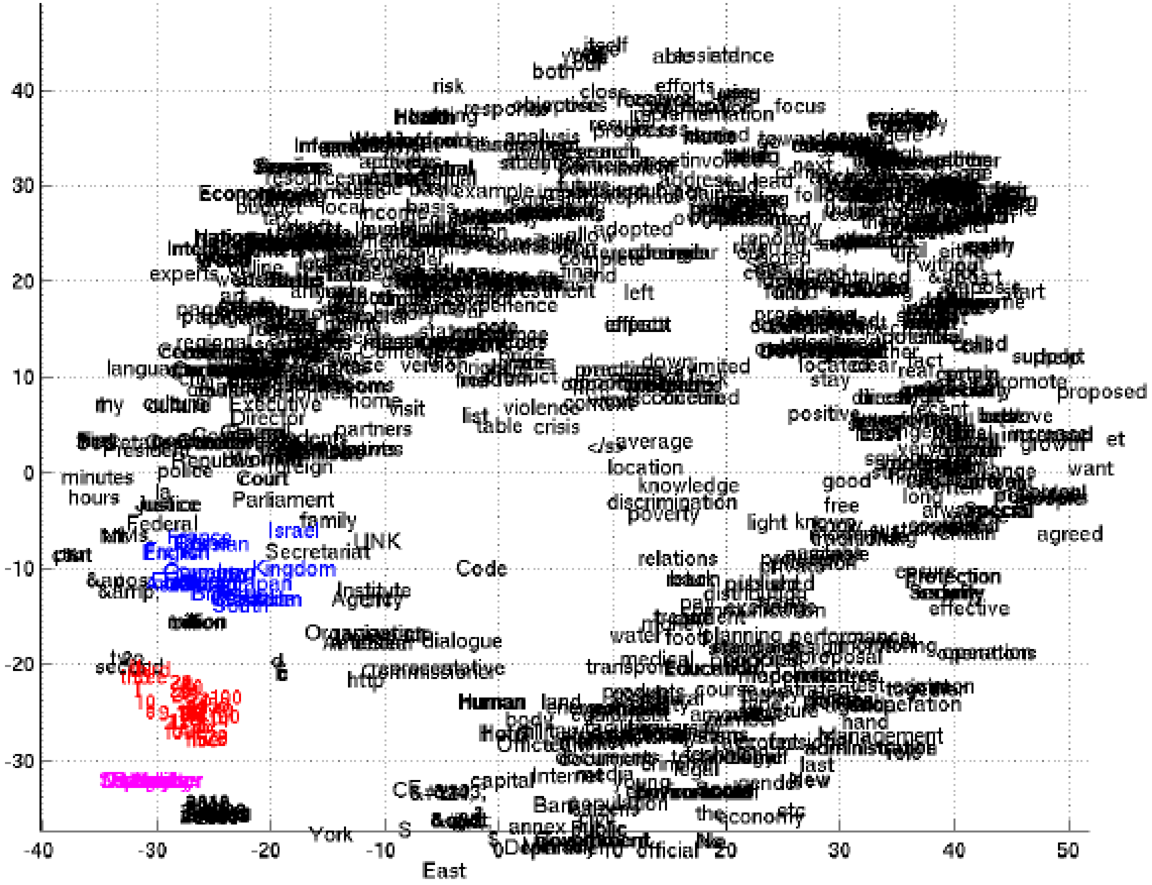


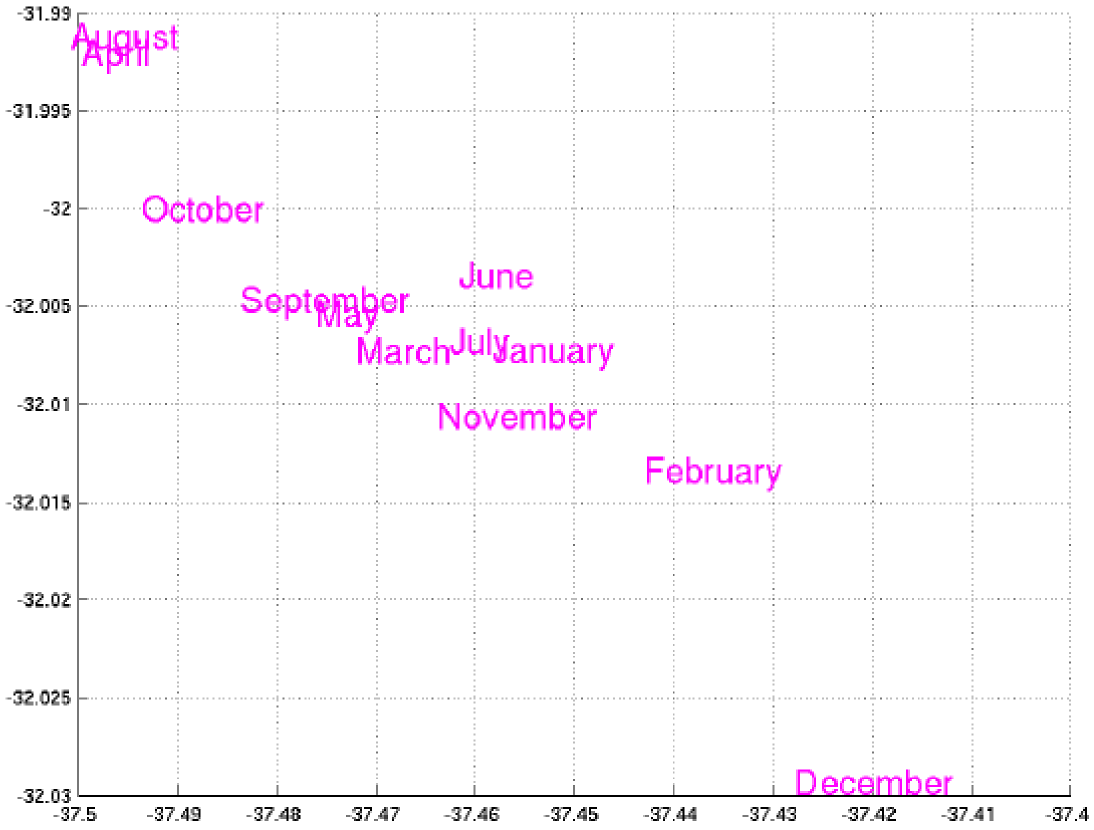


Part 9.

Appendix

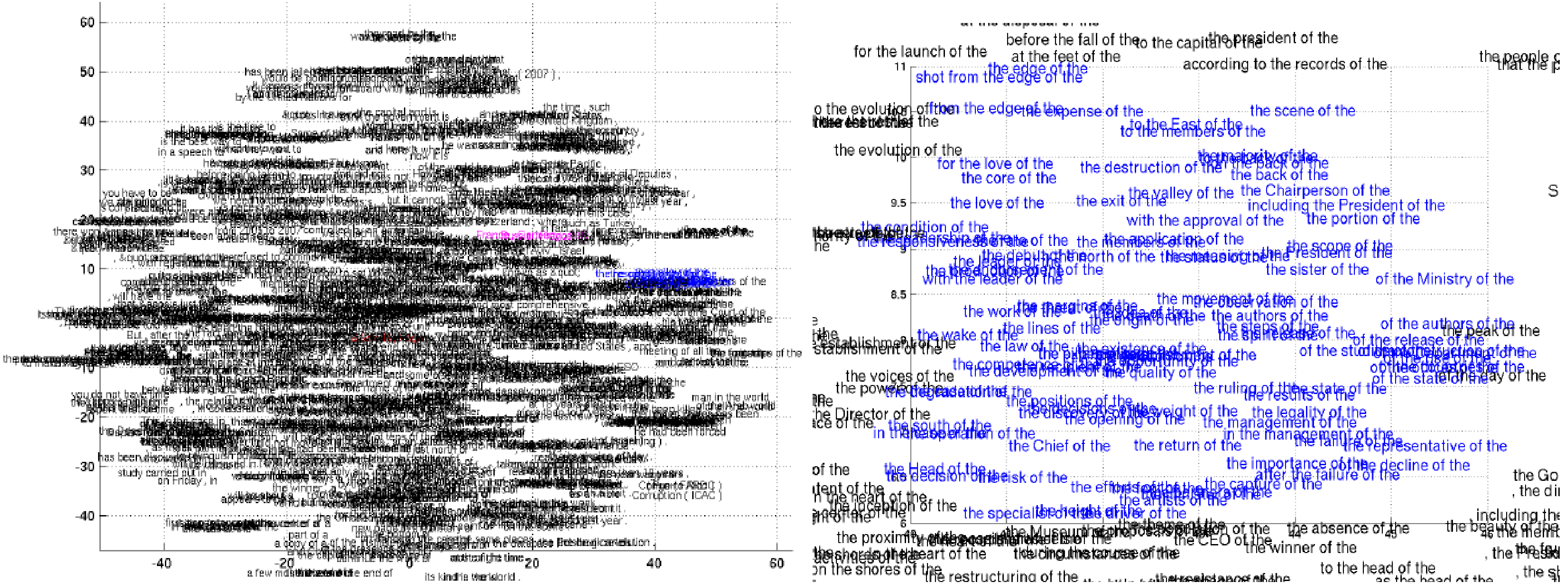
- **2–D embedding of the learned word representation**
  - The left one shows the full embedding space, while the other figures show the zoomed-in view of specific regions (color–coded)





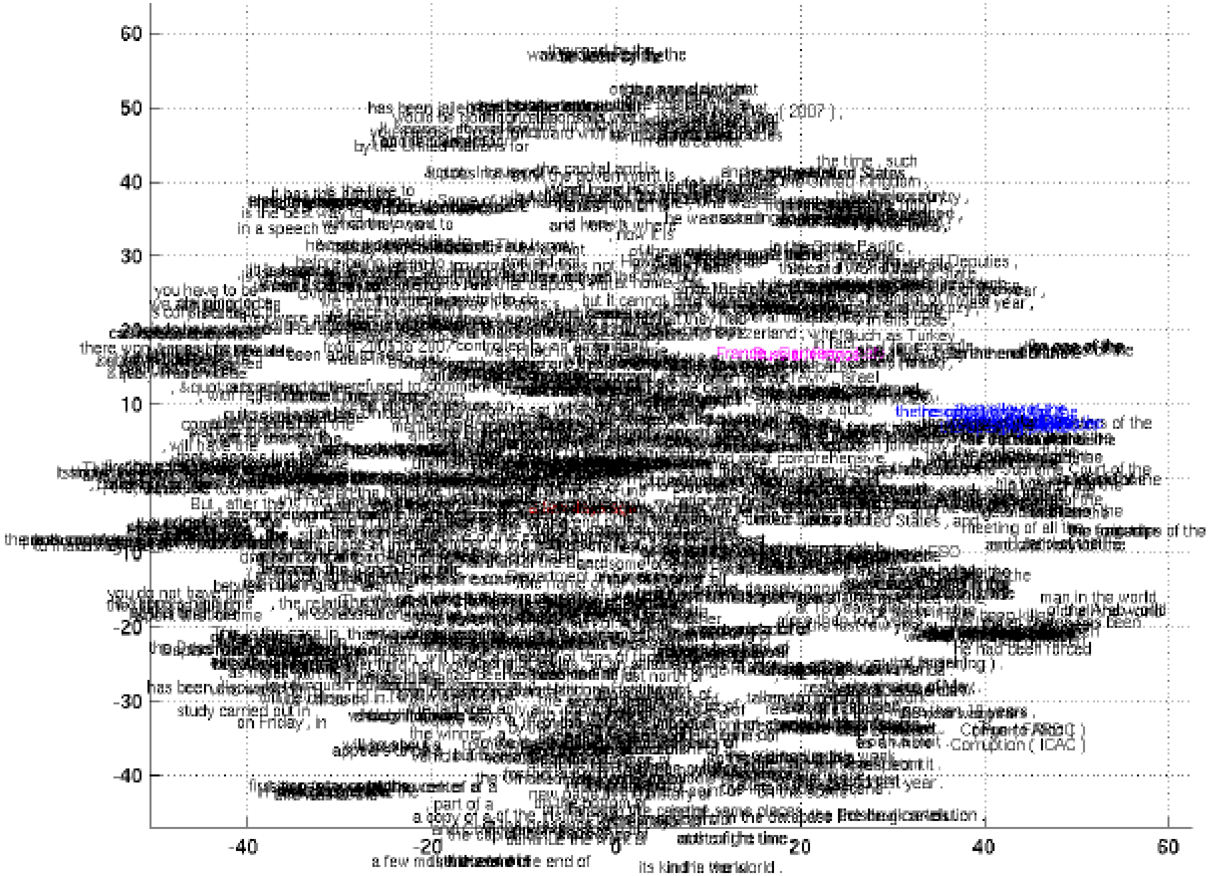
Appendix

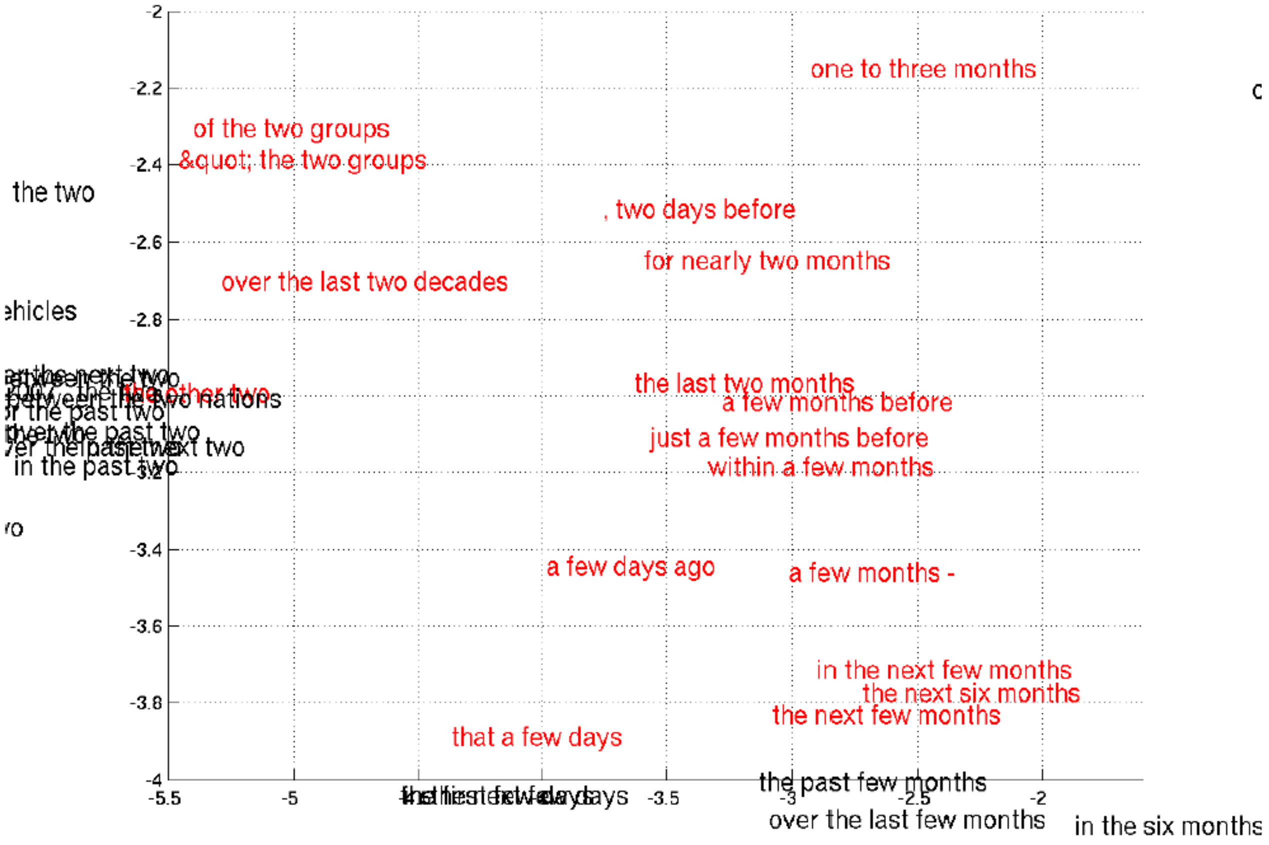
- **2–D embedding of the learned phrase representation**
  - The left one one shows the full representation space (1000 randomly selected points), while the other figures show the zoomed-in view of specific regions (color–coded)



Appendix

- **2–D embedding of the learned phrase representation**
  - The left one one shows the full representation space (1000 randomly selected points), while the other figures show the zoomed-in view of specific regions (color–coded)





Appendix

- **2–D embedding of the learned phrase representation**
  - The left one one shows the full representation space (1000 randomly selected points), while the other figures show the zoomed-in view of specific regions (color–coded)