Bilingual Word Embeddings and Unsupervised SMT

Viktor Hangya
hangyav@cis.lmu.de

CIS, LMU Munich

2023

Outline

Word Embeddings

- Vector representation
- Learning word embeddings

Bilingual Word Embeddings

- Motivation
- Overview of approaches
- Mapping
- Orthogonal mapping
- Unsupervised training

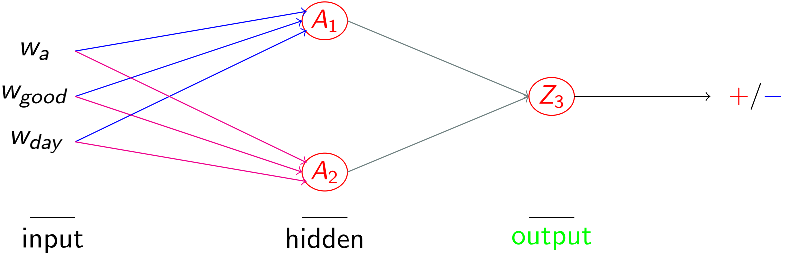
Unsupervised Statistical MT

- MT with bilingual word embeddings
- Integrating embeddings into Moses

Word representation

How do we represent words in neural networks?

- One-hot vector:
  - $w_{good} = [1, 0, 0, 0, ..., 0]$
  - $w_{great} = [0, 1, 0, 0, ..., 0]$
  - $w_{day} = [0, 0, 1, 0, ..., 0]$
  - $w_{a} = [0, 0, 0, 1, ..., 0]$



Word representation

How do we represent words in neural networks?

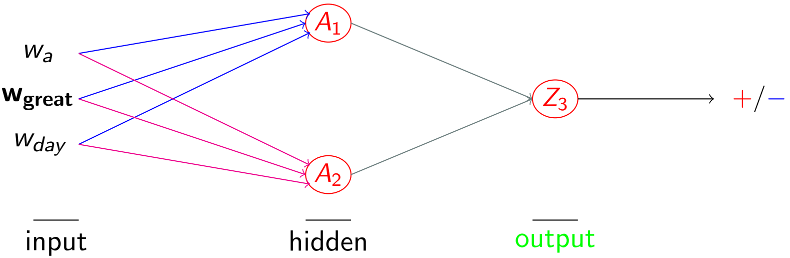
- One-hot vector:
  - $w_{good} \;= [1, 0, 0, 0, ..., 0]$
  - $w_{great} = [0, 1, 0, 0, ..., 0]$
  - $w_{day} \;\;= [0, 0, 1, 0, ..., 0]$
  - $w_{a} \;\;\;\;= [0, 0, 0, 1, ..., 0]$

- Does not carry word similarity information
- This would be better
  - $w_{good} \;= [\mathbf{1},\ \mathbf{0}, 0, 0, ..., 0]$
  - $w_{great} = [\mathbf{1},\ \mathbf{0}, 0, 0, ..., 0]$
  - $w_{day} \;\;= [0,\ 0, 1, 0, ..., 0]$
  - $w_{a} \;\;\;\;= [0,\ 0, 0, 1, ..., 0]$

Word representation

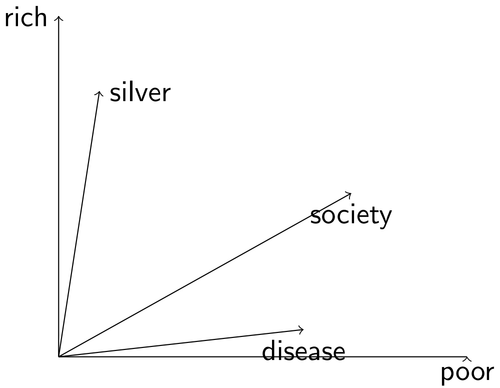
How do we represent words in neural networks?

- One-hot vector:
  - $w_{good} = [\mathbf{1},\ \mathbf{0}, 0, 0, \ldots, 0]$
  - $w_{great} = [\mathbf{1},\ \mathbf{0}, 0, 0, \ldots, 0]$
  - $w_{day} = [0,\ 0, 1, 0, \ldots, 0]$
  - $w_{a} = [0,\ 0, 0, 1, \ldots, 0]$



Word Embeddings

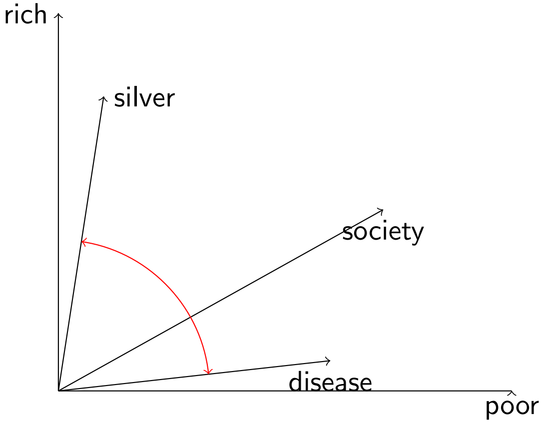
- Representation of words in vector space
- Learn a low dimensional vector representations
  - typically *dimensions* $= 300$
  - $w_{good} = \left[0.234, 0.001, -0.456, 0.000, ..., -0.938\right]$



Word Embeddings

- Similar words are close to each other
  $\rightarrow$ Similarity is the cosine of the angle between two word vectors

$$cosine(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$



Word Embeddings

- Vector representation
- Learning word embeddings

Bilingual Word Embeddings

- Motivation
- Overview of approaches
- Mapping
- Orthogonal mapping
- Unsupervised training

Unsupervised Statistical MT

- MT with bilingual word embeddings
- Integrating embeddings into Moses

Learning word embeddings

Distributional similarity: the meaning of words can be understood from their context

I drink **water** *before bed.* or *I drink* **milk** *before bed.*

Count-based methods:

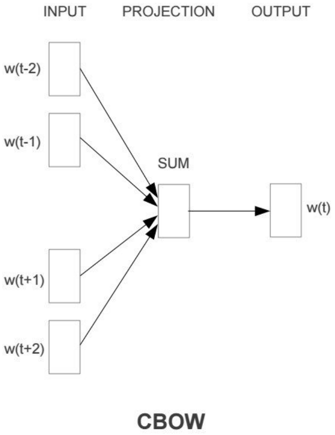
- Compute co-occurrence statistics
- Map sparse high-dimensional vectors to small dense representation
- Matrix factorization approaches: SVD

Neural networks:

- Predict a word from its neighbors
  - **Word2Vec: CBOW and skipgram** Mikolov et al. (2013a)
- Language Modeling Task
  - ELMo, BERT Peters et al. (2018); Devlin et al. (2018)

Learning word embeddings with Continuous Bag-Of-Words

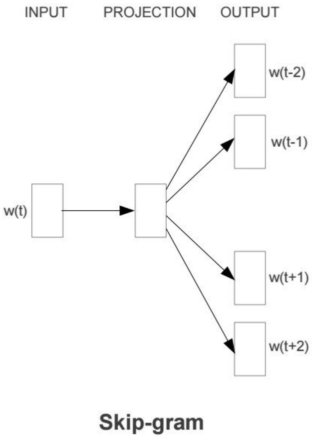
Training example:... *CEO of BMW was fired ...*



Mikolov et al. (2013a)

Learning word embeddings with skip-gram

Training example: ... *CEO of BMW was fired ...*



Mikolov et al. (2013a)

Word embedding quality

- Semantic similarity
  - Correlation of cosine similarities with given word pair similarity scores
  - Example: *SimLex-999*
    - coast – shore: 9.00
    - clothes – closet 1.96
- Downstream tasks
  - Embeddings as features in neural networks
  - Example:
    - Machine translation (BLEU)
    - sentiment analysis (accuracy)
    - etc.

Word Embeddings

- Vector representation
- Learning word embeddings

Bilingual Word Embeddings

- Motivation
- Overview of approaches
- Mapping
- Orthogonal mapping
- Unsupervised training

Unsupervised Statistical MT

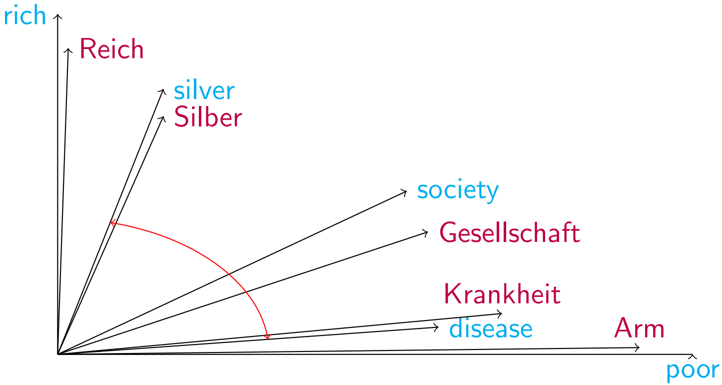
- MT with bilingual word embeddings
- Integrating embeddings into Moses

Bilingual Word Spaces

Representation of words in two languages in same semantic space:

→ Similar words are close to each other

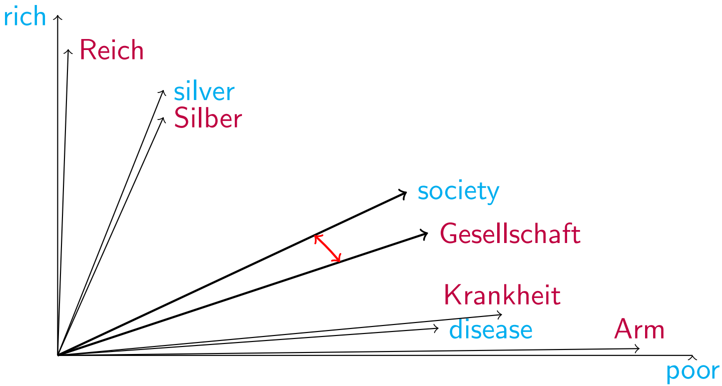
→ Given by cosine



Translating words

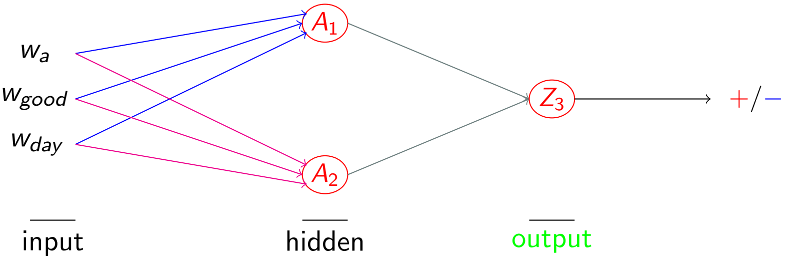
Translating word using cosine similarity

- society → Gesellschaft



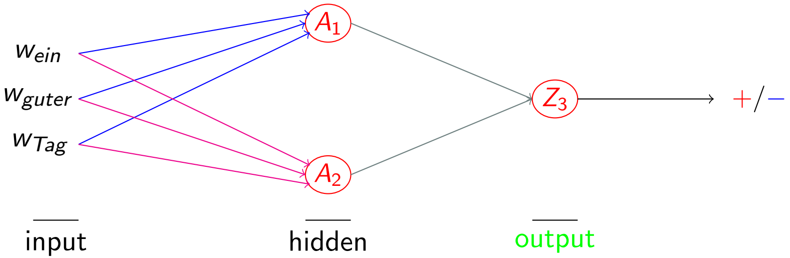
Cross-lingual transfer learning

- Train on English
  - $w_{good} \quad = \left[0.23, 0.01, -0.45, 0.00, ..., 0.93\right]$
  - $w_{day} \quad = \left[-0.76, 0.98, 0.23, 0.74, ..., 0.01\right]$
  - $w_{a} \quad = \left[0.54, -0.39, 0.28, 0.79, ..., 0.42\right]$



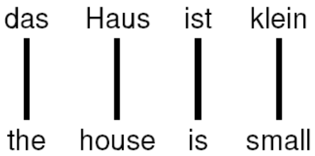
Cross-lingual transfer learning

- Classify German
  - $w_{good} = [0.23, 0.01, -0.45, 0.00, ..., 0.93] \approx w_{guter}$
  - $w_{day} = [-0.76, 0.98, 0.23, 0.74, ..., 0.01] \approx w_{Tag}$
  - $w_a = [0.54, -0.39, 0.28, 0.79, ..., 0.42] \approx w_{ein}$



Learning Bilingual Word Embeddings

- Required bilingual training signal:
  - Parallel sentences
    - ⋆ Hermann and Blunsom (2014), Gouws et al. (2015), Gouws and Søgaard (2015), Duong et al. (2016)



Learning Bilingual Word Embeddings

- Required bilingual training signal:
  - Document-aligned data
    - ★ Vulic and Moens (2015); Vulic and Korhonen (2016)

Dog

From Wikipedia, the free encyclopedia

This article is about the domestic dog. For related species known as "dogs", see Canidae. For other uses, see Dog (disambiguation).

The **dog** (*Canis familiaris* when considered a distinct species or *Canis lupus familiaris* when considered a subspecies of the wolf)[5] is a member of the genus *Canis* (canines), which forms part of the wolf-like canids,[6] and is the most widely abundant terrestrial carnivore.[7][8][9][10][11] The dog and the extant gray wolf are sister taxa[12][13][14] as modern wolves are not closely related to the wolves that were first domesticated,[13][14] which implies that the direct ancestor of the dog is extinct.[15] The dog was the first species to be domesticated,[14][16] and has been selectively bred over millennia for various behaviors, sensory capabilities, and physical attributes.[17]

Their long association with humans has led dogs to be uniquely attuned to human behavior[18] and they are able to thrive on a starch-rich diet that would be inadequate for other canids.[19] Dogs vary widely in shape, size and colors.[20] They perform many roles for humans, such as hunting, herding, pulling loads, protection, assisting police and military, companionship and, more recently, aiding disabled people and therapeutic roles. This influence on human society has given them the sobriquet of "man's best friend".

Domestic dogs

Temporal range: At least 14,200 years ago – present[2]

Haushund

Hund ist eine Weiterleitung auf diesen Artikel. Weitere Bedeutungen sind unter Hund (Begriffsklärung) aufgeführt.

Der **Haushund** (*Canis lupus familiaris*) ist ein Haustier und wird als Heim- und Nutztier gehalten. Seine wilde Stammform ist der Wolf, dem er als Unterart zugeordnet wird. Wann die Domestizierung stattfand, ist umstritten; wissenschaftliche Schätzungen variieren zwischen 15.000 und 100.000 Jahren v. u. Z.

Im engeren Sinn bezeichnet man als Haushund die Hunde, die überwiegend im Haus gehalten werden, und kennzeichnet damit also eine Haltungsform. Historisch wurde ein Hund, der zur Bewachung des Hauses gehalten wird, als Haushund bezeichnet.[1] Eine weitere Verwendung des Begriffs ist die Einschränkung auf sozialisierte (Haus-)Hunde, also Hunde, die an das Zusammenleben mit Menschen in der menschlichen Gesellschaft gewöhnt und an dieses angepasst sind. Damit wird der Haushund abgegrenzt gegen wild lebende, verwilderte oder streunende Hunde, die zwar auch domestiziert, aber nicht sozialisiert sind.[2]

Der Dingo ist ebenfalls ein Haushund, wird jedoch provisorisch als eigenständige Unterart des Wolfes geführt.[3]

Haushund

Learning Bilingual Word Embeddings

- Required bilingual training signal:
  - Monolingual data and a seed dictionary
    - Mikolov et al. (2013b); Faruqui and Dyer (2014); Lazaridou et al. (2015)

```
dog   -- Hund
apple -- Apfel
100   -- 100
...   -- ...
```

Learning Bilingual Word Embeddings

- Required bilingual training signal:
  - Monolingual data only
    - Conneau et al. (2017); Artetxe et al. (2018a)

Learning Bilingual Word Embeddings

- Approaches:
  - Mapping
    - Step 1: build source language embeddings
    - Step 2: build target embeddings
    - Step 3: map them to a shared space
  - Joint training
    - Step 1: build both source and target embeddings in one step

Word Embeddings

- Vector representation
- Learning word embeddings

Bilingual Word Embeddings

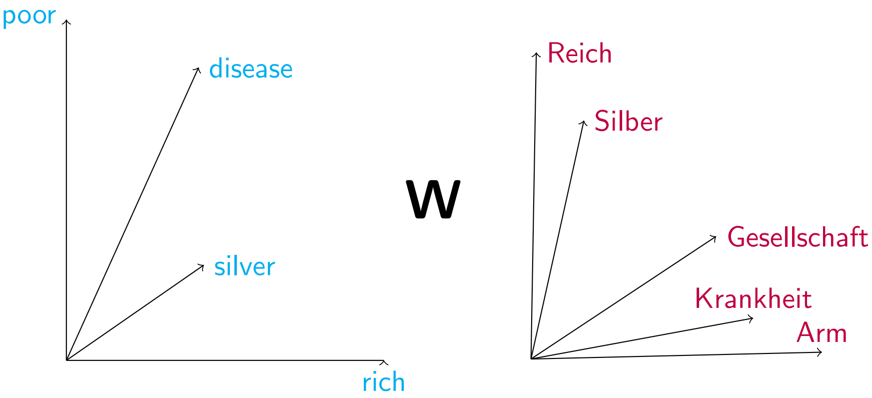
- Motivation
- Overview of approaches
- Mapping
- Orthogonal mapping
- Unsupervised training

Unsupervised Statistical MT

- MT with bilingual word embeddings
- Integrating embeddings into Moses

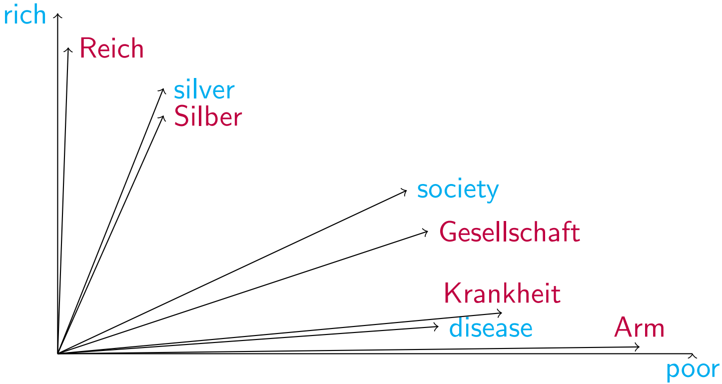
Mapping with seed dictionary

- Learn monolingual word embeddings
- Learn a linear mapping W



Mapping with seed dictionary

- Project source words into target space



Mapping with seed dictionary

1. Train monolingual word embeddings (Word2vec) in English
   - Need English monolingual data
2. Train monolingual word embeddings (Word2vec) in German
   - Need German monolingual data
3. Learn mapping **W** using a seed dictionary
   - Need a list of 5000 English words and their translation

Learning W by minimizing Euclidean distance

Regression (Mikolov et al. (2013b))

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \sum_{\mathbf{i}}^{\mathbf{n}} || \mathbf{W}\mathbf{x_i} - \mathbf{y_i} ||^{\mathbf{2}}$$

x_i : **embedding** of i-th source (English) word in the seed dictionary.

y_i : **embedding** of i-th target (German) word in the seed dictionary.

Learning W by minimizing Euclidean distance

Regression (Mikolov et al. (2013b))

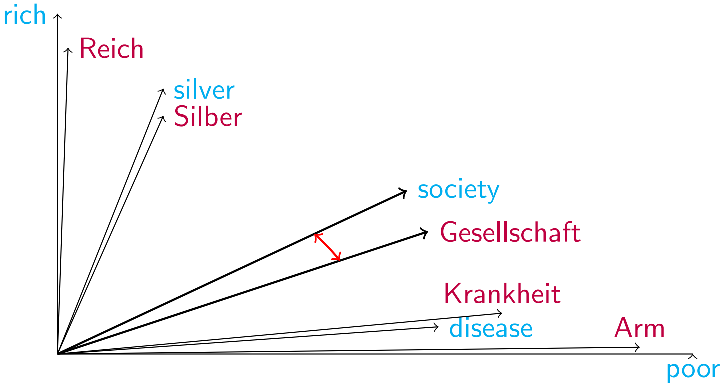
$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \sum_{\mathbf{i}}^{\mathbf{n}} \| \mathbf{W}\mathbf{x_i} - \mathbf{y_i} \|^2$$

- For each pair (x_i, y_i) in the seed dictionary:
  - Predict projection $\mathbf{y*}$ by computing $\mathbf{Wx_i}$
  - Compute squared error between $\mathbf{y*}$ and $\mathbf{y_i}$
- Find $\mathbf{W}$ such that squared error over training set is minimal

Bilingual dictionary induction

- Task to evaluate bilingual word embeddings intrinsically
- Given a set of source words, find the corresponding translations:
  - Given society, find its vector in the BWE space
  - Retrieve the German word whose vector is the most similar (cosine)

$$cosine(\mathbf{W}x, y)$$



Bilingual dictionary induction

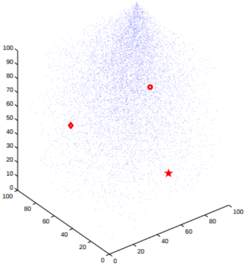
- Evaluation: *precision@n*
  - is the correct translation in the n most similar translations?
  - dog → Katze, Hunde, **Hund**, Giraffe, Maus

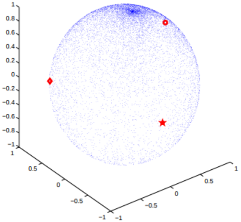
	English to Italian			Italian to English		
	P@1	P@5	P@10	P@1	P@5	P@10
Methods with cross-lingual supervision (WaCky)						
Mikolov et al. (2013b) †	33.8	48.3	53.9	24.9	41.0	47.4

Conneau et al. (2017)

Vector normalization

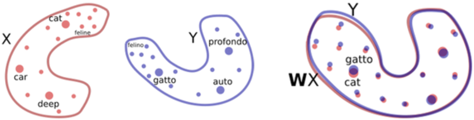
- Mismatch:
  - Training (squared-error): $\mathbf{W}^* = \arg\min_{\mathbf{W}} \sum_{\mathbf{i}}^{\mathbf{n}} || \mathbf{W}\mathbf{x_i} - \mathbf{y_i} ||^{\mathbf{2}}$
  - Test: $cosine(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$
- Normalize vectors to length 1: $x = \frac{x}{\|x\|}$
  - $\mathbf{W}^* = \arg\max_{\mathbf{W}} \sum_{\mathbf{i}}^{\mathbf{n}} (\mathbf{W}\mathbf{x_i})^\top \mathbf{y_i}$



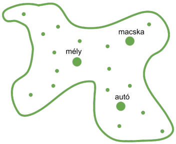


Isomorphism

- Isomorphism (approximate)
  - If languages convey similar information in similar contexts their monolingual embeddings should be isomorphic.
  - it is only true to some extent for each language pair
- restrictions on the mapping (**W**)
  - e.g. allow rotation only
  - preserves word similarities in the monolingual spaces



(Conneau et al., 2017)



Procrustes problem

- Orthogonality constraint on $\mathbf{W}$
- Xing et al. (2015)

$$\mathbf{W}^* = UV^\top \qquad U\Sigma V^\top = SVD(YX^\top)$$

X : contains the embedding of the i^{th} source word (x_i) in the seed dictionary in row i.

Y : contains the embedding of the i^{th} target word (y_i) in the seed dictionary in row i.

Bilingual dictionary induction

	English to Italian			Italian to English		
	P@1	P@5	P@10	P@1	P@5	P@10
Methods with cross-lingual supervision (WaCky)						
Mikolov et al. (2013b) †	33.8	48.3	53.9	24.9	41.0	47.4
Procrustes - CSLS	44.9	61.8	66.6	38.5	57.2	63.0

Conneau et al. (2017)

* Cross-Domain Similarity Local Scaling (CSLS): cosine alternative Conneau et al. (2017)

Unsupervised mapping

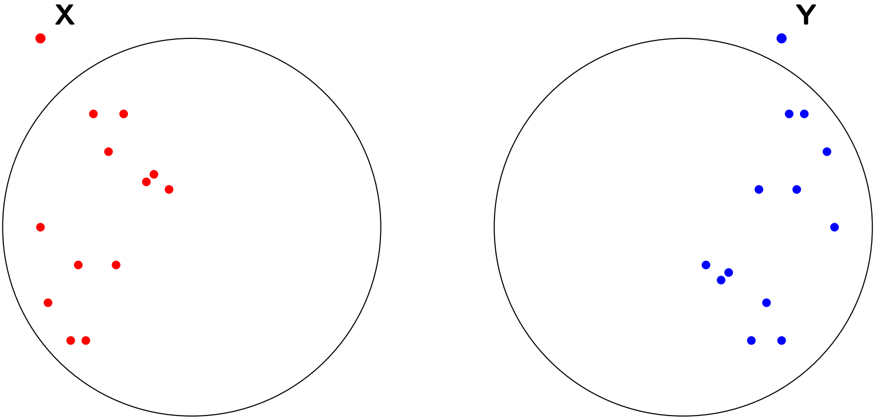
- Low-resource languages
  - No seed dictionary is available for training
  - only monolingual corpora for both languages

0. Generate an initial seed dictionary automatically
1. Learn mapping
2. Induce a better dictionary
3. Goto 1. until convergence

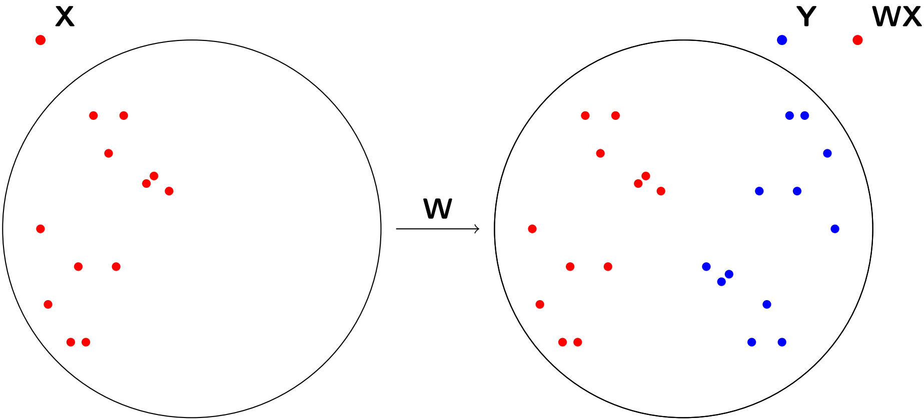
Adversarial training

- Conneau et al. (2017)
- Two player game
  - **Discriminator**: discriminate mapped source language vectors $\mathbf{W}x_i$ from target y_i
    - x_i and y_i are the embeddings of any word in the vocabulary (we have no seed dictionary)
  - **Generator**: generate the mapping $\mathbf{W}$ such that the discriminator fails
- Iterative process: both make a step after each other

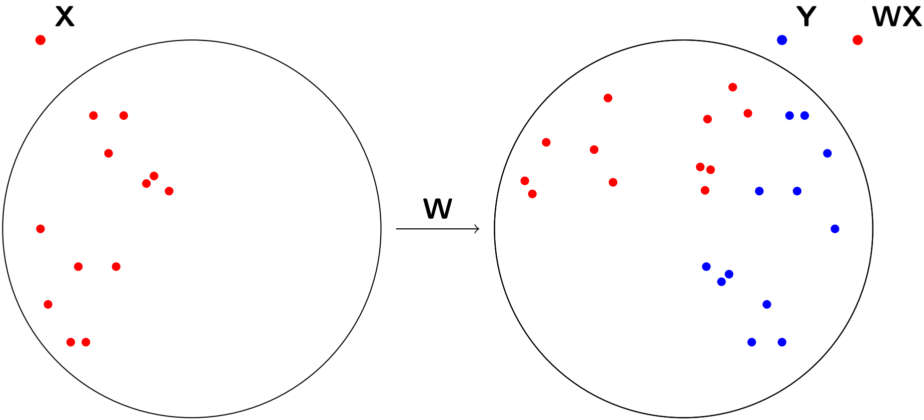
Adversarial training



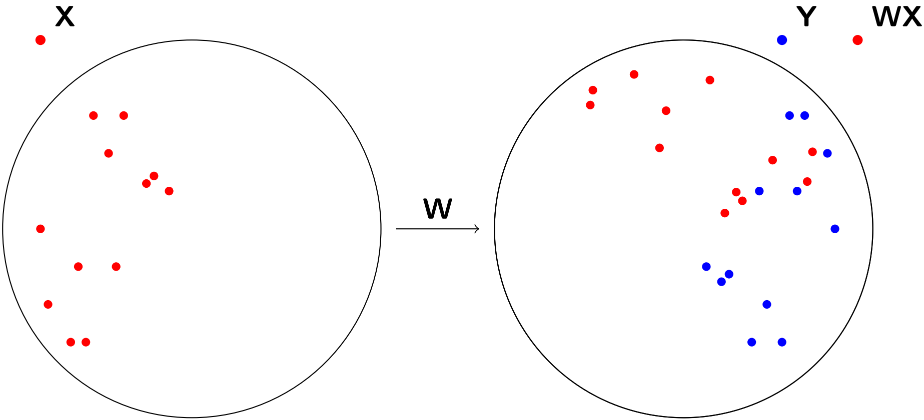
Adversarial training



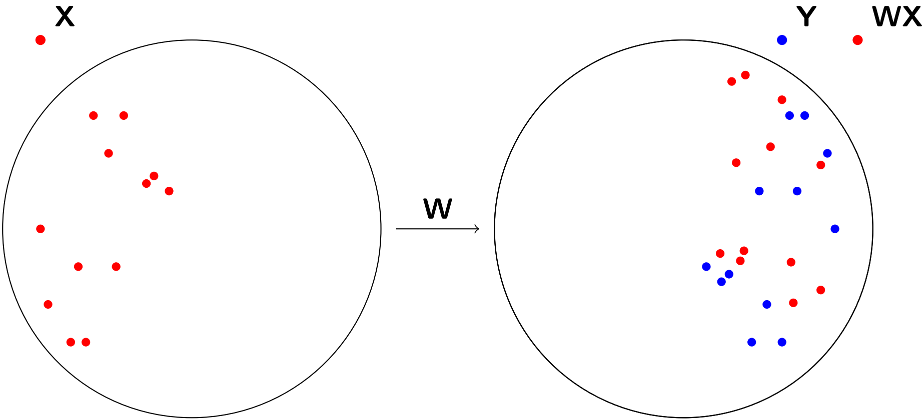
Adversarial training



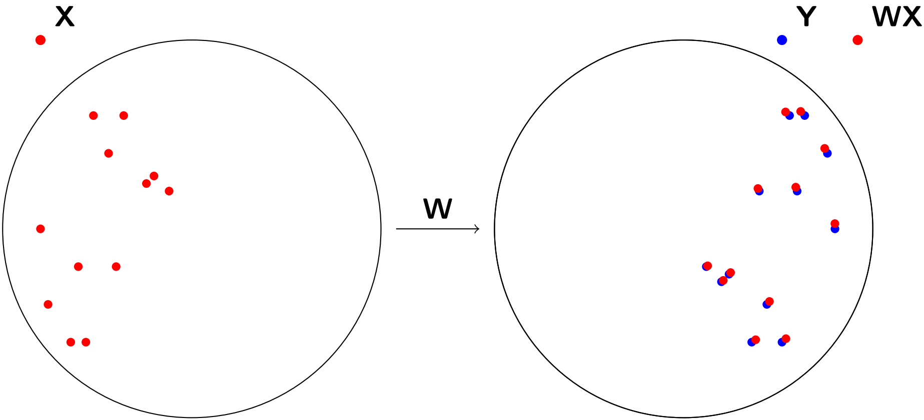
Adversarial training



Adversarial training



Adversarial training



Adversarial training

- Discriminator
  - Feed forward network: probability of vector x is the embedding of a source word $P_{\theta_D}(source = 1|x)$

$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}\left(\text{source} = 1\middle|Wx_i\right) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}\left(\text{source} = 0\middle|y_i\right)$$

Adversarial training

- Discriminator
  - Feed forward network: probability of vector x is the embedding of a source word $P_{\theta_D}(source = 1|x)$

$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}\left(\text{source} = 1\middle|Wx_i\right) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}\left(\text{source} = 0\middle|y_i\right)$$

- Generator

$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}\left(\text{source} = 0\middle|Wx_i\right) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}\left(\text{source} = 1\middle|y_i\right)$$

Adversarial training

- Discriminator
  - Feed forward network: probability of vector x is the embedding of a source word $P_{\theta_D}(source = 1|x)$

$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}\left(\text{source} = 1\middle|Wx_i\right) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}\left(\text{source} = 0\middle|y_i\right)$$

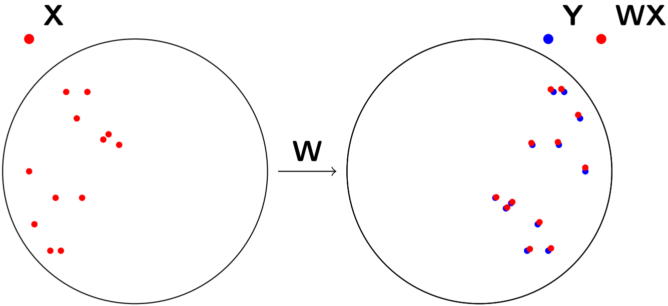
- Generator

$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}\left(\text{source} = 0\middle|Wx_i\right) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}\left(\text{source} = 1\middle|y_i\right)$$

- Minimize both losses with gradient descent

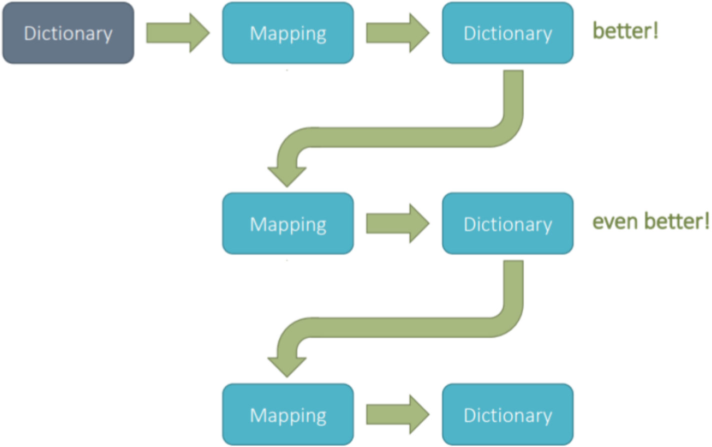
Initial dictionary

- The iterative process gives us an initial W
- Noisy but works well for frequent words
- Translate frequent words with bilingual dictionary induction
- Use this dictionary for orthogonal mapping



Iterative refinement

- Increase the quality and size of the dictionary in each step



Ruder et al. (2019)

Bilingual dictionary induction

	English to Italian			Italian to English		
	P@1	P@5	P@10	P@1	P@5	P@10
Methods with cross-lingual supervision (WaCky)						
Mikolov et al. (2013b) †	33.8	48.3	53.9	24.9	41.0	47.4
Procrustes - CSLS	44.9	61.8	66.6	38.5	57.2	63.0
Methods without cross-lingual supervision (WaCky)						
Adv - Refine - CSLS	45.1	60.7	65.1	38.3	57.8	62.8

Conneau et al. (2017)

Bilingual dictionary induction

- Results are lower for distant language pairs
- Isomorphism is weaker

	Unsupervised (Adversarial)	**Supervised** (Identical)
EN-ES	81.89	**82.62**
EN-ET	00.00	**31.45**
EN-FI	00.09	**28.01**
EN-EL	00.07	**42.96**
EN-HU	45.06	**46.56**
EN-PL	46.83	**52.63**
EN-TR	32.71	**39.22**

Søgaard et al. (2018)

Word Embeddings

- Vector representation
- Learning word embeddings

Bilingual Word Embeddings

- Motivation
- Overview of approaches
- Mapping
- Orthogonal mapping
- Unsupervised training

Unsupervised Statistical MT

- MT with bilingual word embeddings
- Integrating embeddings into Moses

Unsupervised MT

- Most MT systems need a large amount of parallel sentences (few millions)
- Low-resource languages lack this resource
- We use only monolingual data to build an MT system
- Idea:
  - Start simple: word translation
  - Extend the initial system to sentences
- We don't expect to beat supervised systems but for many languages this is the best we can do

Word-by-word translation

- Translate each word in a sentence independently of the others
- Bilingual dictionary induction
  - unsupervised bilingual word embeddings
  - cosine similarity of words

Word-by-word translation

- Translate each word in a sentence independently of the others
- Bilingual dictionary induction
  - unsupervised bilingual word embeddings
  - cosine similarity of words

Der	Himmel	ist	blau	.
The	sky	is	blue	.

Word-by-word translation

- Translate each word in a sentence independently of the others
- Bilingual dictionary induction
  - unsupervised bilingual word embeddings
  - cosine similarity of words

Der	Himmel	ist	blau	.
The	sky	is	blue	.

Word-by-word translation

- Translate each word in a sentence independently of the others
- Bilingual dictionary induction
  - unsupervised bilingual word embeddings
  - cosine similarity of words

Der	**Himmel**	ist	blau	.
The	**sky**	is	blue	.

Word-by-word translation

- Translate each word in a sentence independently of the others
- Bilingual dictionary induction
  - unsupervised bilingual word embeddings
  - cosine similarity of words

Der	Himmel	**ist**	blau	.
The	sky	**is**	blue	.

Word-by-word translation

- Translate each word in a sentence independently of the others
- Bilingual dictionary induction
  - unsupervised bilingual word embeddings
  - cosine similarity of words

Der	Himmel	ist	**blau**	.
The	sky	is	**blue**	.

Word-by-word translation

- Translate each word in a sentence independently of the others
- Bilingual dictionary induction
  - unsupervised bilingual word embeddings
  - cosine similarity of words

Der	Himmel	ist	blau	.
The	sky	is	blue	.

Results

	en-fr	fr-en	de-en	en-de
Supervised	27.97	26.13	25.61	21.33
word-by-word	6.28	10.09	10.77	7.06

Lampple et al. (2018)

Problems

- Translating compound words

Der	Himmel	ist	**dunkelblau**	**.**
The	sky	is	~~**dark**~~	**.**
The	sky	is	~~**blue**~~	**.**
The	sky	is	***dark blue***	**.**

Problems

- Translating multi word expressions

Fix	und	fertig
~~Fixed~~	~~and~~	~~ready~~

Exhausted

Phrase embeddings

- One embedding for n-grams
  - $w_{dark\_blue} = [0.23, 0.01, -0.45, 0.00, ..., 0.93]$

1. look for frequently co-occurring n-grams in the monolingual corpus (Mikolov et al. (2013c))

$$score(w_i, w_j) = \frac{count(w_i, w_j)}{count(w_i) \times count(w_j)}$$

2. Concatenate n-grams in the corpus
   - The sky is **dark_blue** .

3. Train monolingual embeddings as before
4. Mapping of monolingual spaces

Phrase translation

- Translate tokens with cosine
- Tokens can be phrases as well

Der Himmel ist **dunkelblau** .

- $cosine(dunkelblau, dark) = 0.7$
- $cosine(dunkelblau, blue) = 0.75$
- $cosine(dunkelblau, \mathbf{dark\_blue}) = 0.83$

fix und fertig

- $cosine(fix, fixed) = 0.68$
- $cosine(und, and) = 0.8$
- $cosine(fertig, ready) = 0.7$
- $cosine(\mathbf{fix\_und\_fertig}, exhausted) = 0.95$

Word Embeddings

- Vector representation
- Learning word embeddings

Bilingual Word Embeddings

- Motivation
- Overview of approaches
- Mapping
- Orthogonal mapping
- Unsupervised training

Unsupervised Statistical MT

- MT with bilingual word embeddings
- Integrating embeddings into Moses

Problems

- Translating words independently leads to problems:
  - Fluency/word order

Ich	denke	dass	der	Himmel	blau	ist
I	think	that	the	sky	~~blue~~	~~is~~
I	think	that	the	sky	**is**	**blue**

Problems

- Translating words independently leads to problems:
  - Multi sense words and morphology

Ich	sitze	auf	der	Bank
~~I~~	~~sit~~	on	the	~~bank~~
I'm	**sitting**	on	the	**bench**

Log-linear model

- Feature functions:
  - Phrase-table → adequacy
    - ★ Generate n (100) most probable translations for each source word/phrase with bilingual dictionary induction
    - → different morphological variations and senses

$$\phi(f|e) = \frac{e^{cos(e,f)}}{\sum_{\hat{f}} e^{cos(e,\hat{f})}}$$

  - Language model → fluency
    - → right sense and morphology of words given the source sentence
    - → right order of words
    - ★ Train the same way as for supervised SMT
    - ★ using the same monolingual corpus as for the embeddings

Tuning weights

- Finding the right feature weights needs parallel data
- Back-translation:
  0. Build systems for both directions: $S_{L_1 \rightarrow L_2}$ and $S_{L_2 \rightarrow L_1}$
     - ⋆ uniform feature weights
  1. Generate synthetic parallel data
     - ⋆ $D_{L_2 \rightarrow L_1}$: (back-)translate L_1 monolingual data with $S_{L_1 \rightarrow L_2}$
     - ⋆ $D_{L_1 \rightarrow L_2}$: (back-)translate L_2 monolingual data with $S_{L_2 \rightarrow L_1}$
     - ⋆ Source language is noisy but target is not
  2. Optimize weight with MERT
     - ⋆ $D_{L_2 \rightarrow L_1}$ for $S_{L_2 \rightarrow L_1}$
     - ⋆ $D_{L_1 \rightarrow L_2}$ for $S_{L_1 \rightarrow L_2}$
  3. Goto 1. until convergence

Iterative back-translation

- Only the weights are tuned in the previous step
- Build SMT systems from scratch using synthetic parallel data
  - word align sentences
  - build phrase tables
  - etc.

- Weights of $S_{L_1 \to L_2}$ and $S_{L_2 \to L_1}$ are now tuned
- Back-translation:
  1. Generate synthetic parallel data
     - $D_{L_2 \to L_1}$: (back-)translate L_1 monolingual data with $S_{L_1 \to L_2}$
     - $D_{L_1 \to L_2}$: (back-)translate L_2 monolingual data with $S_{L_2 \to L_1}$
  2. Build **supervised** SMT systems from scratch
     - $D_{L_1 \to L_2}$ for $S_{L_1 \to L_2}$
     - $D_{L_2 \to L_1}$ for $S_{L_2 \to L_1}$
  3. Goto 1. until convergence

Results

	en-fr	fr-en	de-en	en-de
Supervised	27.97	26.13	25.61	21.33
word-by-word	6.28	10.09	10.77	7.06

Lample et al. (2018)

	FR-EN	**EN-FR**	**DE-EN**	**EN-DE**
Proposed system	**25.87**	**26.22**	**17.43**	**14.08**

Artetxe et al. (2018b)

Summary

- Word embeddings
  - Similar words have similar word vectors
  - Training them with *word2vec* using monolingual data
- Bilingual word embeddings
  - Mapping monolingual embeddings to shared space using seed dictionary
  - Unsupervised mapping with adversarial initialization
- Unsupervised SMT
  - Word-by-word translation with unsupervised bilingual embeddings
  - Integrating bilingual dictionary induction to SMT

Thank you !

References I

Artetxe, M., Labaka, G., and Agirre, E. (2018a). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789—-798.

Artetxe, M., Labaka, G., and Agirre, E. (2018b). Unsupervised Statistical Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632—-3642.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

References II

Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T. (2016). Learning crosslingual word embeddings without bilingual corpora. In *Proc. EMNLP*.

Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proc. EACL*.

Gouws, S., Bengio, Y., and Corrado, G. (2015). Bilbowa: Fast bilingual distributed representations without word alignments. In *Proc. ICML*.

Gouws, S. and Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In *Proc. NAACL*.

Hermann, K. M. and Blunsom, P. (2014). Multilingual models for compositional distributed semantics. In *Proc. ACL*, pages 58–68, Baltimore, Maryland. Association for Computational Linguistics.

Lample, G., Denoyer, L., and Ranzato, M. (2018). Unsupervised Machine Translation Using Monolingual Corpora Only. In *Proceedings of the International Conference on Learning Representations*, pages 1–12.

References III

Lazaridou, A., Dinu, G., and Baroni, M. (2015). Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proc. ACL*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013c). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

References IV

Ruder, S., Søgaard, A., and Vulić, I. (2019). Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38.

Søgaard, A., Ruder, S., and Vulić, I. (2018). On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788.

Vulic, I. and Korhonen, A. (2016). On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *Proc. ACL*, pages 247–257.

Vulic, I. and Moens, M. (2015). Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proc. ACL*.

Xing, C., Wang, D., Liu, C., and Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proc. NAACL*.