Cross-Lingual Word Embeddings
Unsupervised Machine Translation

Cristina España-Bonet
DFKI GmbH

Low-Resource NLP:
Multilinguality and Machine Translation
Webinar Series — Session II
29th June 2021

Session II (& III?): Unsupervised Neural Machine Translation

Ingredients

1. Data

- Monolingual corpora

2. Initialisation

- Cross-lingual embeddings
- Deep MLM pretraining

3. Training

SMT and/or NMT

- Denoising autoencoder
- Backtranslation

Session II (& III?): Unsupervised Neural Machine Translation

Ingredients for Today

1. **Data**

   - **Monolingual corpora**

2. **Initialisation**

   - **Cross-lingual embeddings**
   - Deep MLM pretraining

3. **Training**

   SMT and/or **NMT**

   - **Denoising autoencoder**
   - **Backtranslation**

Session II (& III?): Unsupervised Neural Machine Translation

Ingredients for the Next Time

1. Data

- Monolingual corpora

2. Initialisation

- Cross-lingual embeddings
- **Deep MLM pretraining**

3. Training

SMT and/or **NMT**

- **Denoising autoencoder**
- **Backtranslation**

Session II (& III?): Unsupervised Neural Machine Translation

Ingredients as Homework

1. **Data**
   - Monolingual corpora
2. **Initialisation**
   - Cross-lingual embeddings
   - Deep MLM pretraining
3. **Training**

   SMT and/or NMT

   - Denoising autoencoder
   - Backtranslation

Session II (& III?): Unsupervised Neural Machine Translation

Inspiration & Borrowing

Cross-lingual embeddings

Sebastian Ruder, Anders Søgaard and Ivan Vulić

ACL 2019 tutorial. (`https://tinyurl.com/xlingual`)

Unsupervised machine translation

Mikel Artetxe

PhD thesis and related presentations. (`shorturl.at/wBELP`)

Rui Wang and Hai Zhao

EACL 2021 tutorial. Advances and Challenges in Unsupervised Neural Machine Translation (joint CLWE+UMT and multilingual UMT)

(`https://wangruinlp.github.io/unmt`)

Session II

Outline

1. Recap through the Examples of Session I

2. Word Embeddings
   - ~~Basics~~
   - ~~Frequency and Prediction-based Embeddings~~
   - Cross-lingual Embeddings

3. Unsupervised Machine Translation

Recap through the Examples of Session I

What's the Meaning of Low-Resource?

Definition (for us!). *A low-resource setting is a scenario where standard NLP techniques are not usable (low/null performance).*

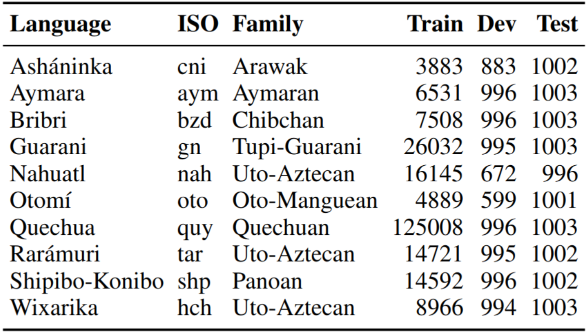
I talk about **low-resource setting** because

- Task dependent
  - speech recognition vs. machine translation vs. PoS tagging
- Language (complexity) dependent
  - English vs. Hungarian
- Domain dependent!
  - English text generation: sport vs. corona in March 2020
- Author dependent!

Recap through the Examples of Session I

Example: What is Low-Resource Machine Translation?

AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas (Mager et al. 2021)



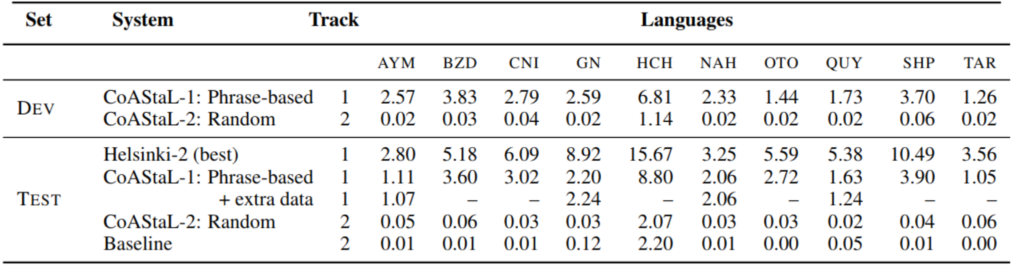
Language	ISO	Family	Train	Dev	Test
Asháninka	cni	Arawak	3883	883	1002
Aymara	aym	Aymaran	6531	996	1003
Bribri	bzd	Chibchan	7508	996	1003
Guarani	gn	Tupi-Guarani	26032	995	1003
Nahuatl	nah	Uto-Aztecan	16145	672	996
Otomí	oto	Oto-Manguean	4889	599	1001
Quechua	quy	Quechuan	125008	996	1003
Rarámuri	tar	Uto-Aztecan	14721	995	1002
Shipibo-Konibo	shp	Panoan	14592	996	1002
Wixarika	hch	Uto-Aztecan	8966	994	1003

Recap through the Examples of Session I

Example: What is Low-Resource Machine Translation?

AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas (Bollmann et al. 2021)

BLEU scores



Set	System	Track	Languages									
			AYM	BZD	CNI	GN	HCH	NAH	OTO	QUY	SHP	TAR
DEV	CoAStaL-1: Phrase-based	1	2.57	3.83	2.79	2.59	6.81	2.33	1.44	1.73	3.70	1.26
	CoAStaL-2: Random	2	0.02	0.03	0.04	0.02	1.14	0.02	0.02	0.02	0.06	0.02
TEST	Helsinki-2 (best)	1	2.80	5.18	6.09	8.92	15.67	3.25	5.59	5.38	10.49	3.56
	CoAStaL-1: Phrase-based	1	1.11	3.60	3.02	2.20	8.80	2.06	2.72	1.63	3.90	1.05
	+ extra data	1	1.07	–	–	2.24	–	2.06	–	1.24	–	–
	CoAStaL-2: Random	2	0.05	0.06	0.03	0.03	2.07	0.03	0.03	0.02	0.04	0.06
	Baseline	2	0.01	0.01	0.01	0.12	2.20	0.01	0.00	0.05	0.01	0.00

Recap through the Examples of Session I

Main Approaches to LR-NLP

1. Data enrichment
   - Data collection
   - Data augmentation
2. General machine learning
   - Unsupervised learning
   - Weak supervision
   - Transfer learning
3. Multilinguality and/or multimodality
4. Specialised architectures

Recap through the Examples of Session I

Example: Basic Low-Resource NLP. MT Yorùbá–English (Adelani et al., 2021)

Model (tested on MENYO-20k)	*en***2***yo*	*yo***2***en*
JW300+Bible **baseline**	8.1±0.2	10.8±0.3
+Transfer learning **domain adaptation**	12.3±0.3	13.2±0.3
JW300+Bible+MENYO-20k **domain adaptation**	10.9±0.3	14.0±0.3
+Transfer learning **domain adaptation**	**12.4±0.3**	14.6±0.3
+ Backtranslation **data augmentation**	12.0±0.3	**18.2±0.4**

Recap through the Examples of Session I

Example: Basic Low-Resource NLP. MT Yorùbá–English (Adelani et al., 2021)

Model (tested on MENYO-20k)	*en***2***yo*	*yo***2***en*
JW300+Bible **baseline**	8.1±0.2	10.8±0.3
+Transfer learning **domain adaptation**	12.3±0.3	13.2±0.3
JW300+Bible+MENYO-20k **domain adaptation**	10.9±0.3	14.0±0.3
+Transfer learning **domain adaptation**	**12.4±0.3**	14.6±0.3
+ Backtranslation **data augmentation**	12.0±0.3	**18.2±0.4**
mT5-base+Transfer learning **pretraining** **task adaptation**	11.5±0.3	16.3±0.4

Recap through the Examples of Session I

Example: Basic Low-Resource NLP. MT Yorùbá–English (Adelani et al., 2021)

Model (tested on Menyo-20k)	*en*2*yo*	*yo*2*en*
JW300+Bible **baseline**	8.1±0.2	10.8±0.3
+Transfer learning **domain adaptation**	12.3±0.3	13.2±0.3
JW300+Bible+Menyo-20k **domain adaptation**	10.9±0.3	14.0±0.3
+Transfer learning **domain adaptation**	**12.4±0.3**	14.6±0.3
+ Backtranslation **data augmentation**	12.0±0.3	**18.2±0.4**
mT5-base+Transfer learning **pretraining** **task adaptation**	11.5±0.3	16.3±0.4
Google GMNMT **multilingual**	3.7±0.2	**22.4±0.5**
Facebook M2M-100 **multilingual**	3.3±0.2	4.6±0.3
OPUS-MT **bilingual**	–	5.9±0.2

Session II

Outline

1 Recap through the Examples of Session I

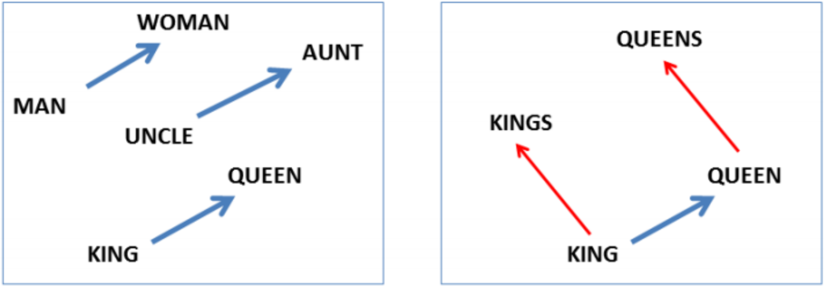
2 Word Embeddings
- ~~Basics~~
- ~~Frequency and Prediction-based Embeddings~~
- Cross-lingual Embeddings

3 Unsupervised Machine Translation

Monolingual Embeddings (Recap!)

What we all Know about Embeddings

King - Man + Woman = Queen



(Mikolov et al., NAACL HLT, 2013)

Monolingual Embeddings (Recap!)

Types of Embeddings

Frequency-based Embeddings

- Term frequency, TF-IDF, co-occurrence matrix

Prediction-based Embeddings

- GloVe, skip-gram, CBoW, etc.

Basic Unit

- word (word2vec, GloVe, etc.), *n*-gram (fastText), character (CWE)

Monolingual Embeddings (Recap!)

Evaluation of Word Embeddings

Extrinsic Methods

Performance in a downstream NLP task

- Text classification, NER, PoS tagging, etc.

"Intrinsic" Methods

Correlation with human judgments on words relations

- Word semantic similarity (WordSim, SemEval, SimVerb, etc.),
- Word analogy (SemEval, WordRep, MSR, etc.)

Unfortunately, methods do **not correlate** among themselves!

Monolingual Embeddings (Recap!)

In the Low-Resource Setting...

- Few data affects the quality of the embeddings
- Noise in data affects the quality of the embeddings
- Domain mistmatch between training and task affects the performance of the embeddings
- The choice of the *correct* architecture might be more critical
- Languages other than English are difficult to evaluate

Monolingual Embeddings (Recap!)

Example in LR: Yorùbá and Twi (Alabi et al., 2020)

Description	Source URL	#tokens	Status	C1	C2	C3
Yorùbá						
Lagos-NWU corpus	github.com/Niger-Volta-LTI	24,868	clean	✓	✓	✓
Alákọ̀wé	alakoweyoruba.wordpress.com	24,092	clean	✓	✓	✓
Ọ̀rọ̀ Yorùbá	oroyoruba.blogspot.com	16,232	clean	✓	✓	✓
Èdè Yorùbá Rẹwà	deskgram.cc/edeyorubarewa	4,464	clean	✓	✓	✓
Doctrine $ Covenants	github.com/Niger-Volta-LTI	20,447	clean	✓	✓	✓
Bible	www.bible.com	819,101	clean	✓	✓	✓
GlobalVoices	yo.globalvoices.org	24,617	clean	✓	✓	✓
Jehovah's Witness	www.jw.org/yo	170,203	clean	✓	✓	✓
Ìrìnkèrindò nínú igbó elégbèje	manual	56,434	clean	✓	✓	✓
Igbó Olódùmarè	manual	62,125	clean	✓	✓	✓
JW300	opus.nlpl.eu/JW300.php	10,558,055	clean	✗	✗	✓
YorùbáTweets	twitter.com/yobamoodua	153,716	clean	✓	✓	✓
BBC Yorùbá	bbc.com/yoruba	330,490	noisy	✗	✓	✓
Voice of Nigeria Yorùbánews	von.gov.ng/yoruba	380,252	noisy	✗	✗	✓
Wikipedia	dumps.wikimedia.org/yowiki	129,075	noisy	✗	✗	✓
Twi						
Bible	www.bible.com	661,229	clean	✓	✓	✓
Jehovah's Witness	www.jw.org/tw	1,847,875	noisy	✗	✗	✓
Wikipedia	dumps.wikimedia.org/twwiki	5,820	noisy	✗	✓	✓
JW300	opus.nlpl.eu/JW300.php	13,630,514	noisy	✗	✗	✓

Monolingual Embeddings (Recap!)

Example in LR: Yorùbá and Twi (Alabi et al., 2020)

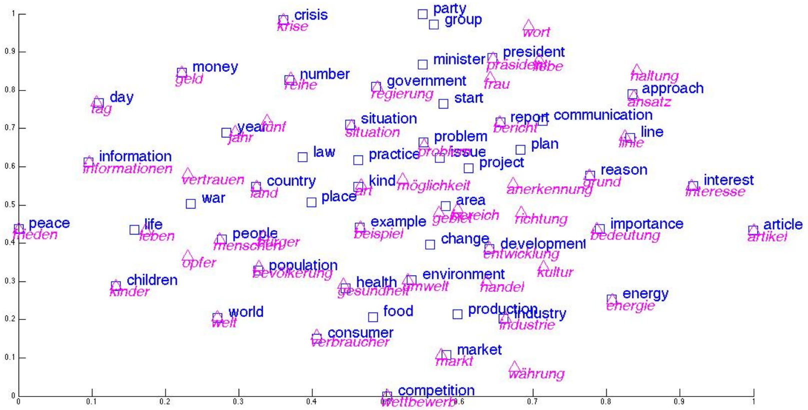
- FastText embeddings, intrinsic eval on wordsim-353 (manually translated)

	Twi		Yorùbá	
Model	**Vocab Size**	**Spearman ρ**	**Vocab Size**	**Spearman ρ**
F1: Pre-trained Model (Wiki)	935	0.143	21,730	0.136
F2: Pre-trained Model (Common Crawl & Wiki)	NA	NA	151,125	0.073
C1: Curated *Small* Dataset (Clean text)	9,923	0.354	12,268	0.322
C2: Curated *Small* Dataset (Clean + some noisy text)	18,494	**0.388**	17,492	0.302
C3: Curated *Large* Dataset (All Clean + Noisy texts)	47,134	0.386	44,560	**0.391**

Monolingual Embeddings (Recap!)

Nice Properties beyond King - Man + Woman = Queen

(Luong, Pham & Manning, NAACL, 2015)



Barnes-Hut-SNE visualisation of bilingual embeddings German/English

Monolingual Embeddings (Recap!)

Bilingualism, Nice Property!

How do we achieve this bilingualism?

Cross-lingual embeddings,
bilingual embeddings,
multi-lingual embeddings

Session II

Outline

1. Recap through the Examples of Session I

2. Word Embeddings
   - ~~Basics~~
   - ~~Frequency and Prediction-based Embeddings~~
   - Cross-lingual Embeddings

3. Unsupervised Machine Translation

Cross-lingual Embeddings

Taxonomy

1. **Supervised**
    - Joint learning
        - Regularization term in the loss function
        - Creating pseudo-bilingual corpora
    - Mapping (post-hoc alignment)
2. **Unsupervised**
    - Mapping with self-learning
    - Mapping with adversarial training

Cross-lingual Embeddings

Whys

Why cross-lingual embeddings?

- Multilingual modeling of meaning
- Support for cross-lingual NLP

Why supervised cross-lingual embeddings?

- Simplicity
- Supervision mostly possible (small dictionaries, common words...)

Why unsupervised cross-lingual embeddings?

- Sometimes outperformed supervised ones!
- Cases without dictionaries

Cross-lingual Embeddings

Summary of Approaches

- The summary is not comprehensive at all (cannot!)
- Selection biased towards understanding unsupervised NMT
- Methods used for low-resource NLP
- Lot of info coming from Sebastian Ruder's blogs and tutorials. Don't miss them!

Supervised Cross-lingual Embeddings

Form of Cross-lingual Supervision

- **Word level**: bilingual dictionaries, word alignments
- **Sentence level**: parallel corpora, sentence aligments
- **Document level**: comparable corpora, document alignments

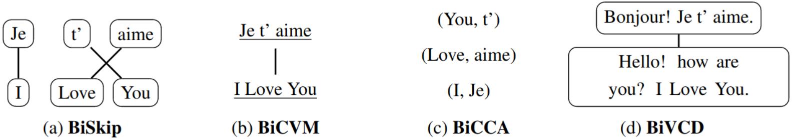
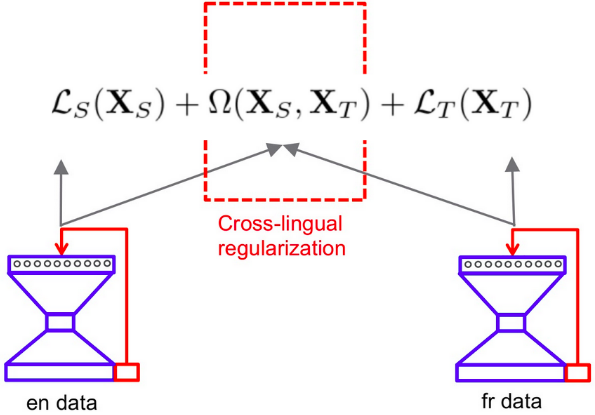


Figure 2: Forms of supervision required by the four models compared in this paper. From left to right, the cost of the supervision required varies from expensive (BiSkip) to cheap (BiVCD). BiSkip requires a parallel corpus annotated with word alignments (Fig. 2a), BiCVM requires a sentence-aligned corpus (Fig. 2b), BiCCA only requires a bilingual lexicon (Fig. 2c) and BiVCD requires comparable documents (Fig. 2d).

(Upadhyay, Faruqui, Dyer & Roth, ACL, 2016)

Supervised Cross-lingual Embeddings

Joint Learning Approaches



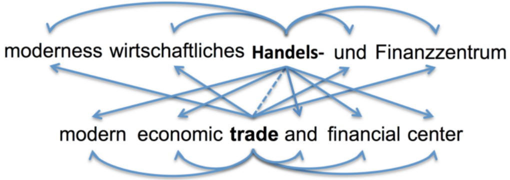
$$\mathcal{L}_S(\mathbf{X}_S) + \Omega(\mathbf{X}_S, \mathbf{X}_T) + \mathcal{L}_T(\mathbf{X}_T)$$

https://tinyurl.com/xlingual

Supervised Cross-lingual Embeddings

Joint Learning Approaches: Bilingual Skipgram

Luong et al., 2015: Bilingual skipgram, direct but expensive

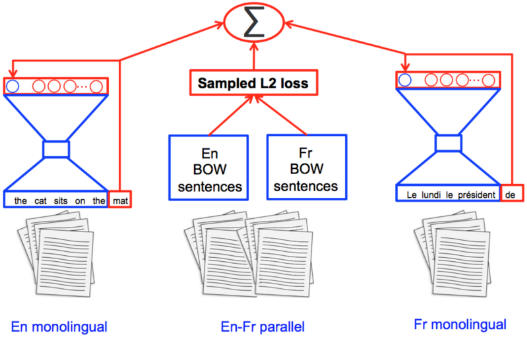

moderness wirtschaftliches Handels- und Finanzzentrum
modern economic trade and financial center

- predict words in the source language **and** predict aligned words in the target language
- parallel corpora + (learned) word aligments

Supervised Cross-lingual Embeddings

Joint Learning Approaches: Bilingual BilBOWA

Guows et al., 2015: Bilingual Bag-of-Words without Word Alignments (**Coulmance et al., 2015**: Trans-gram)

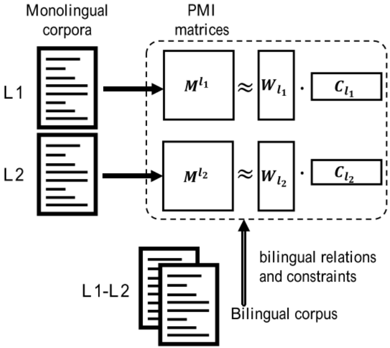


- monolingual skipgram loss
- every word in Source is uniformly aligned to every word in Target
- BilBOWA: minimise distance between the means of the words in the aligned sentences
- Trans-gram: every word in Target as context of every word in Source

Supervised Cross-lingual Embeddings

Joint Learning Approaches: Matrix Co-factorization

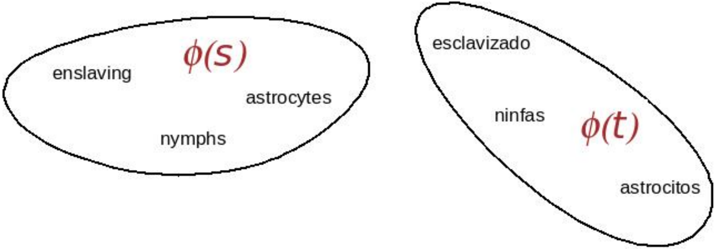
Shi et al., 2015: Joint matrix factorisation


Monolingual corpora
PMI matrices
L1
L2
M^{l_1} ≈ W_{l_1} · C_{l_1}
M^{l_2} ≈ W_{l_2} · C_{l_2}
L1-L2
bilingual relations and constraints
Bilingual corpus

- monolingual GloVe loss
- Ω_1: cross-lingual co-occurrence counts
- Ω_2: minimise the distances of the representations of related words in the two languages weighted by SMT probs
- parallel corpora + (learned) word aligments

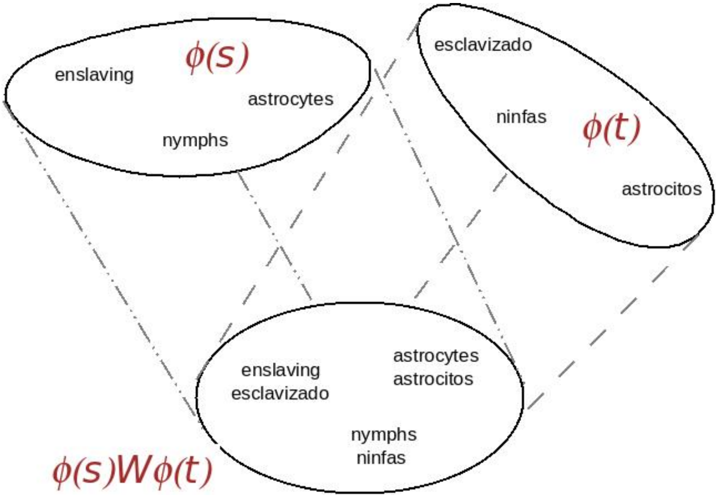
Supervised Cross-lingual Embeddings

Mapping Approaches



Supervised Cross-lingual Embeddings

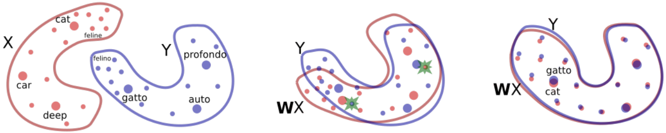
Mapping Approaches



Supervised Cross-lingual Embeddings

Mapping Approaches: Isomorphism (and Other!) Assumption

Spaces should be isomorphic for (linear) mappings to be effective

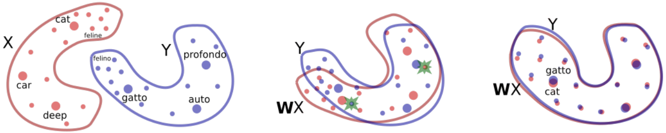


(Figure from *Conneau et al., 2017)*

Supervised Cross-lingual Embeddings

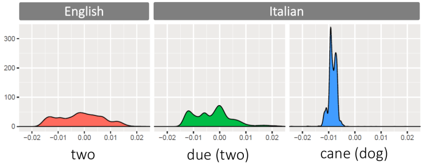
Mapping Approaches: Isomorphism (and Other!) Assumption

Spaces should be isomorphic for (linear) mappings to be effective



(Figure from *Conneau et al., 2017)*

Similarly, similar intra-lingual similarity would be expected



(Figure from *Artetxe et al., 2018)*

Supervised Cross-lingual Embeddings

Mapping Approaches, a bit of History (towards UnsupMT!)

Mikolov et al., 2013: Minimise Euclidean distance

$$W^* = \arg\min_W \parallel Wx_i - y_i \parallel^2, \qquad (x_i, y_i) \text{ pairs in a dictionary}$$

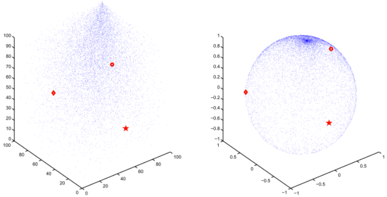
Supervised Cross-lingual Embeddings

Mapping Approaches, a bit of History (towards UnsupMT!)

Mikolov et al., 2013: Minimise Euclidean distance

$$W^* = \arg\min_W \parallel Wx_i - y_i \parallel^2, \qquad (x_i, y_i) \text{ pairs in a dictionary}$$

Xing et al., 2015: Minimise Cosine distance



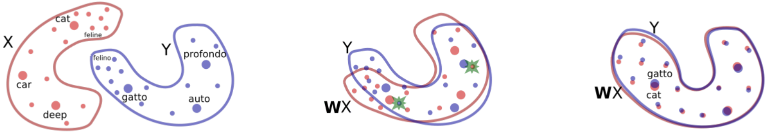
Mismatch between the initial objective function, the distance measure, and the test distance measure

$$W^* = \arg\max_W \cos(Wx_i, y_i)$$

Supervised Cross-lingual Embeddings

Mapping Approaches, a bit of History (towards UnsupMT!)

- The optimisation problem has no closed-form solution
- If W^* is orthogonal, it has a closed-form solution
- Better results when W^* is orthogonal
- Orthogonality preserves monolingual vector space topology



(Conneau et al., 2017)

- If W^* is orthogonal, **Procrustes Problem**

Supervised Cross-lingual Embeddings

Procrustes, the Bandit from Attica

Back into Greece..

https://www.storyboardthat.com/es/storyboards/kaslam/procrustes-2

Supervised Cross-lingual Embeddings

Orthogonal Procrustes Problem

Which is the orthogonal matrix W that most closely maps $X \rightarrow Y$?

$$\arg\min_{W} \|XW - Y\|_F \quad \text{subject to} \quad W^T W = I$$

Supervised Cross-lingual Embeddings

Orthogonal Procrustes Problem

Which is the orthogonal matrix W that most closely maps $X \rightarrow Y$?

$$\arg\min_{W} \|XW - Y\|_F \quad \text{subject to} \quad W^T W = I$$

that is... the optimal rotation and/or reflection (i.e., the optimal orthogonal linear transformation)

Supervised Cross-lingual Embeddings

Orthogonal Procrustes Problem

Which is the orthogonal matrix W that most closely maps $X \to Y$?

$$\arg\min_{W} \|XW - Y\|_F \quad \text{subject to} \quad W^T W = I$$

that is... the optimal rotation and/or reflection (i.e., the optimal orthogonal linear transformation)

Solution: $W = UV^T$ where $X^T Y = M = U\Sigma V^T \Rightarrow \text{SVD}(YX^T)$!

Supervised Cross-lingual Embeddings

Where are we?

1. We have monolingual embeddings
2. We have a (small) dictionary
3. We solve the Procrustes problem to find the projection matrix W
4. Given a word in L1 and W, the equivalent word in L2 can be found by its nearest neighbours according to a similarity measure (cosine?)

Is it all so nice? Almost... the **hubness problem**

Supervised Cross-lingual Embeddings

The Hubness Problem

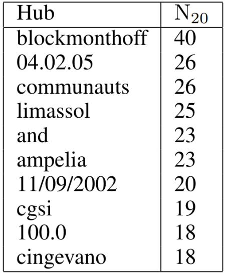
The curse of dimensionality, hubs

In a high-dimensional space, a small set of source vectors (the hubs), appear too frequently in the neighborhood of target vectors

For bilingual WE, some words are close to lots of target words, so they appear in lots of NNs

Example: English → Italian

(Dinu et al., ICLR, 2015)



Hub	N_{20}
blockmonthoff	40
04.02.05	26
communauts	26
limassol	25
and	23
ampelia	23
11/09/2002	20
cgsi	19
100.0	18
cingevano	18

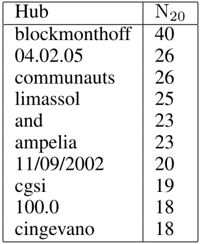
Supervised Cross-lingual Embeddings

The Hubness Problem

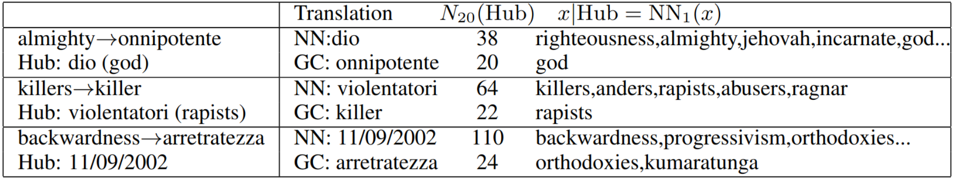
For bilingual WE, some words are close to lots of target words, so they appear in lots of NNs

Example: English → Italian

(Dinu et al., ICLR, 2015)



Hub	N_{20}
blockmonthoff	40
04.02.05	26
communauts	26
limassol	25
and	23
ampelia	23
11/09/2002	20
cgsi	19
100.0	18
cingevano	18



	Translation	N_{20}(Hub)	$x\|\text{Hub} = \text{NN}_1(x)$
almighty→onnipotente Hub: dio (god)	NN:dio GC: onnipotente	38 20	righteousness,almighty,jehovah,incarnate,god... god
killers→killer Hub: violentatori (rapists)	NN: violentatori GC: killer	64 22	killers,anders,rapists,abusers,ragnar rapists
backwardness→arretratezza Hub: 11/09/2002	NN: 11/09/2002 GC: arretratezza	110 24	backwardness,progressivism,orthodoxies... orthodoxies,kumaratunga

Supervised Cross-lingual Embeddings

The Hubness Problem

- Hubs appear in high-dimensional vectors
  - Word embeddings
  - Sentence embeddings (we'll find this later again!)
  - ...
- Different ways to mitigate the problem.
  Relevant for the next systems, rescaling cosine similarity:
  - Margin-based similarity
  - Discounting similarity in dense areas (/,-)

Supervised Cross-lingual Embeddings

Margin-based and Cross-domain Similarity Local Scaling (CSLS)

$$\mathrm{margin}_{\mathrm{CSLS}}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \cos(S_{\mathrm{L1}}, S_{\mathrm{L2}}) - \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 - \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2$$

where $\mathrm{avr}_{\mathrm{kNN}}(X, Y_k) = \sum_{Y \in k\mathrm{NN}(X)} \frac{\cos(X,Y)}{k}$ (average similarity)

Supervised Cross-lingual Embeddings

Margin-based and Cross-domain Similarity Local Scaling (CSLS)

Conneau et al., ICLR, 2018

$$\mathrm{margin}_{\mathrm{CSLS}}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \cos(S_{\mathrm{L1}}, S_{\mathrm{L2}}) - \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 - \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2$$

where $\mathrm{avr}_{\mathrm{kNN}}(X, Y_k) = \sum_{Y \in k\mathrm{NN}(X)} \frac{\cos(X,Y)}{k}$ (average similarity)

Supervised Cross-lingual Embeddings

Margin-based and Cross-domain Similarity Local Scaling (CSLS)

Conneau et al., ICLR, 2018

$$\mathrm{margin}_{\mathrm{CSLS}}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \cos(S_{\mathrm{L1}}, S_{\mathrm{L2}}) - \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 - \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2$$

Artetxe & Schwenk, ACL, 2019

$$\mathrm{margin}_{\mathrm{LASER}}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \frac{\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})}{\mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 + \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2}$$

where $\mathrm{avr}_{\mathrm{kNN}}(X, Y_k) = \sum_{Y \in k\mathrm{NN}(X)} \frac{\cos(X,Y)}{k}$ (average similarity)

Supervised Cross-lingual Embeddings

Where are we?

1. We have monolingual embeddings
2. We have a (small) dictionary
3. We solve the **Procrustes problem** to find the projection matrix W
4. Given a word in L1 and W, the equivalent word in L2 can be found by its nearest neighbours according to a **margin-based similarity** measure

Supervised Cross-lingual Embeddings

Joint learning vs. Mapping

1. **Supervised**
   - **Joint learning**
     - Regularization term in the loss function
     - Creating pseudo-bilingual corpora
   - **Mapping** (post-hoc alignment)
2. **Unsupervised**
   - Mapping with self-learning
   - Mapping with adversarial training

Supervised Cross-lingual Embeddings

Joint learning vs. Mapping

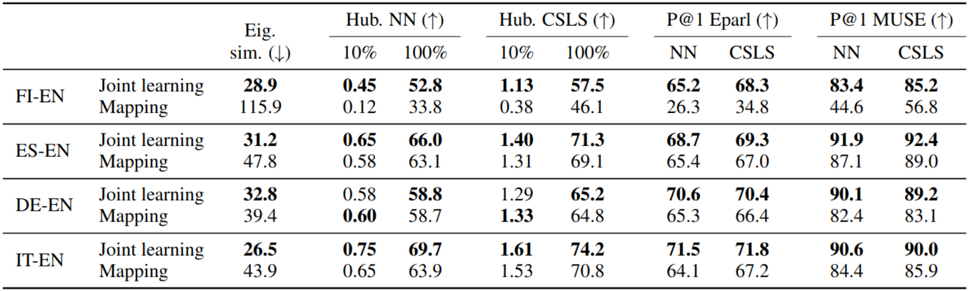
Remember, we rely on the isomorphism assumption of spaces. But,

- separately trained embeddings are not approximately isomorphic in general *Søgaard et al. (2018)*. It depends on
  - the language pair, the comparability of the training corpora, and the parameters of the word embedding algorithms
- the assumption weakens for etymologically distant languages *Patra et al. (2019)*
- embedding spaces in different languages are linearly equivalent only at local regions *Nakashole and Flauger (2018)*
- in the **low-resource setting**, data might not be enough for good monolingual embeddings

Supervised Cross-lingual Embeddings

Joint learning vs. Mapping with Parallel Data, Bilingual Lexicon Induction

Ormanzabal et al., ACL, 2019: Mapping virtues and drawbacks



		Eig.	Hub. NN (↑)		Hub. CSLS (↑)		P@1 Eparl (↑)		P@1 MUSE (↑)	
		sim. (↓)	10%	100%	10%	100%	NN	CSLS	NN	CSLS
FI-EN	Joint learning	**28.9**	**0.45**	**52.8**	**1.13**	**57.5**	**65.2**	**68.3**	**83.4**	**85.2**
	Mapping	115.9	0.12	33.8	0.38	46.1	26.3	34.8	44.6	56.8
ES-EN	Joint learning	**31.2**	**0.65**	**66.0**	**1.40**	**71.3**	**68.7**	**69.3**	**91.9**	**92.4**
	Mapping	47.8	0.58	63.1	1.31	69.1	65.4	67.0	87.1	89.0
DE-EN	Joint learning	**32.8**	0.58	**58.8**	1.29	**65.2**	**70.6**	**70.4**	**90.1**	**89.2**
	Mapping	39.4	**0.60**	58.7	**1.33**	64.8	65.3	66.4	82.4	83.1
IT-EN	Joint learning	**26.5**	**0.75**	**69.7**	**1.61**	**74.2**	**71.5**	**71.8**	**90.6**	**90.0**
	Mapping	43.9	0.65	63.9	1.53	70.8	64.1	67.2	84.4	85.9

Table 1: Evaluation measures for the two cross-lingual embedding approaches. Arrows indicate whether lower (↓) or higher (↑) is better. See text for further details.

Supervised Cross-lingual Embeddings

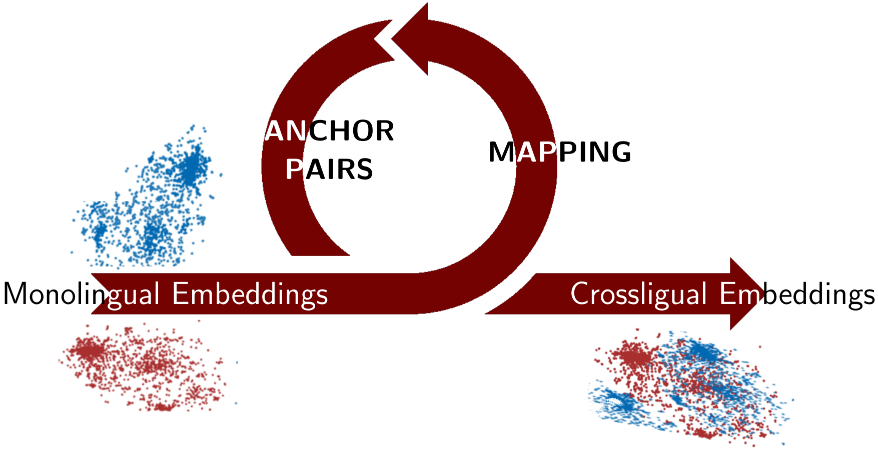
From Supervised Mapping to Unsupervised Self-Learning

1. **Supervised**
   - Joint learning
     - Regularization term in the loss function
     - Creating pseudo-bilingual corpora
   - **Mapping** (post-hoc alignment)

2. **Unsupervised**
   - **Mapping with self-learning**
   - Mapping with adversarial training

Unsupervised Cross-lingual Embeddings

Self-Learning: Simple Idea, Hard Implementation



ANCHOR PAIRS

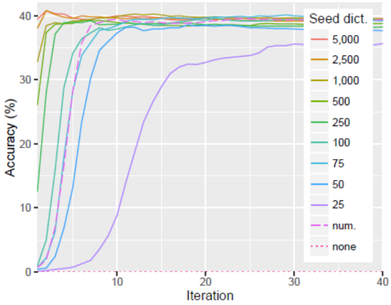
MAPPING

Monolingual Embeddings

Crossligual Embeddings

Unsupervised Cross-lingual Embeddings

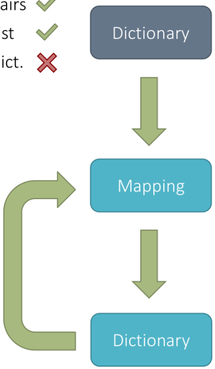
Self-Learning (Mikel Artetxe Slide)



- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

$$W^* = \underset{W \in O(n)}{\arg\min} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

Self-learning



The difference between supervised and unsupervised is the (induction of) the **seed dictionary**

Unsupervised Cross-lingual Embeddings

Self-Learning Basics

1. (Induce —*isomorphism!*) initial seed lexicon $D^{(0)}$

2. Mapping: learn the (linear —*isomorphism!*) projection $W^{(k)}$ with $D^{(k)}$

3. Induce a new dictionary $D^{(k+1)}$ from $XW^{(k)}$

Unsupervised Cross-lingual Embeddings

Self-Learning Basics

1. (Induce —*isomorphism!*) initial seed lexicon $D^{(0)}$
   - Similarity of monolingual similarity distributions
   - Adversarial learning
   - PCA-based similarity
   - Solving optimal transport problem

2. Mapping: learn the (linear —*isomorphism!*) projection $W^{(k)}$ with $D^{(k)}$

3. Induce a new dictionary $D^{(k+1)}$ from $XW^{(k)}$

Unsupervised Cross-lingual Embeddings

Self-Learning Basics

1. (Induce —*isomorphism!*) initial seed lexicon $D^{(0)}$
   - Similarity of monolingual similarity distributions
   - Adversarial learning
   - PCA-based similarity
   - Solving optimal transport problem

2. Mapping: learn the (linear —*isomorphism!*) projection $W^{(k)}$ with $D^{(k)}$
   - Procrustes problem

3. Induce a new dictionary $D^{(k+1)}$ from $XW^{(k)}$
   - Given a word in L1 and W, the equivalent word in L2 can be found by its nearest neighbours according to a margin-based similarity (CSLS) measure

Unsupervised Cross-lingual Embeddings

The Importance of Pre/Post-Processing

Pre-mapping

Normalisation: unit length normalisation, mean centering

Whitening: turning covariance matrices into the identity matrix (unit variance for each dim)

Post-mapping

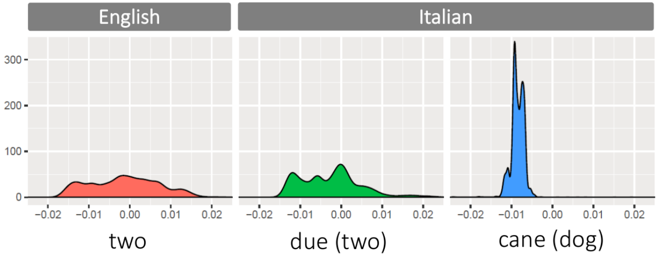
Re-weighting: re-weight each component according to its cross-correlation to increase the relevance of those that best match across languages

De-whitening: restore the original variance in each dimension

Dimensionality reduction: keep only the first n components of the resulting embeddings (and set the rest to 0)

Unsupervised Cross-lingual Embeddings

Lexicon Induction via Heuristics (Artetxe et al., ACL, 2018)



- Words with similar meaning have similar monolingual similarity distributions
- Monolingual similarity: XX^T

Unsupervised Cross-lingual Embeddings

Lexicon Induction via Heuristics (Artetxe et al., ACL, 2018)

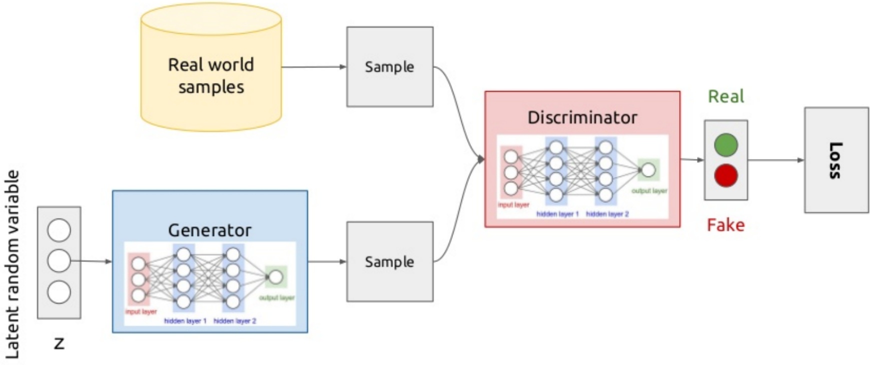
- XX^T dot product between all word combinations in a language.
  Intra-lingual similarity distribution

- Smoothed monolingual similarity distribution:
  $X' = \mathrm{sorted}(\sqrt{XX^T})$ and $Y' = \mathrm{sorted}(\sqrt{YY^T})$

- Dictionary: Nearest neighbours from X' and Y'.
  Similarity between similarities!

Unsupervised Cross-lingual Embeddings

Lexicon Induction via Adversarial Training (Conneau et al., ICLR, 2018)



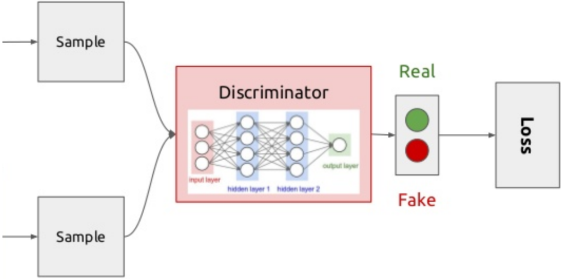
Unsupervised Cross-lingual Embeddings

Lexicon Induction via Adversarial Training (Conneau et al., ICLR, 2018)

Y

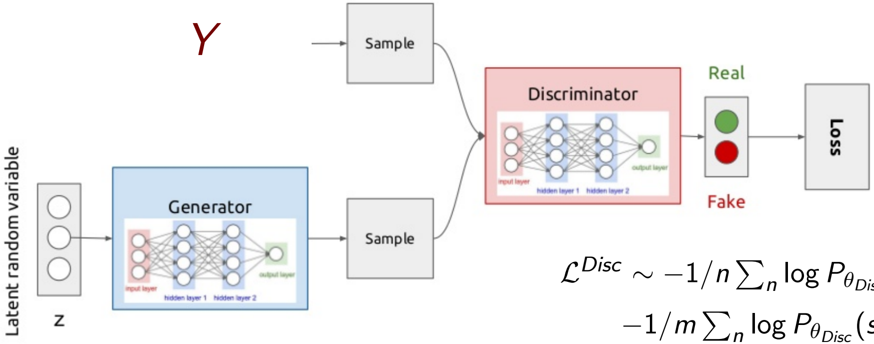
XW

Generator



Unsupervised Cross-lingual Embeddings

Lexicon Induction via Adversarial Training (Conneau et al., ICLR, 2018)



$$\mathcal{L}^{Disc} \sim -1/n \sum_n \log P_{\theta_{Disc}}(src = 1|Wx_i)$$
$$-1/m \sum_n \log P_{\theta_{Disc}}(src = 0|y_i)$$

$$\mathcal{L}^{Gen} \sim -1/n \sum_n \log P_{\theta_{Disc}}(src = 0|Wx_i)$$
$$-1/m \sum_n \log P_{\theta_{Disc}}(src = 1|y_i)$$

Unsupervised Cross-lingual Embeddings

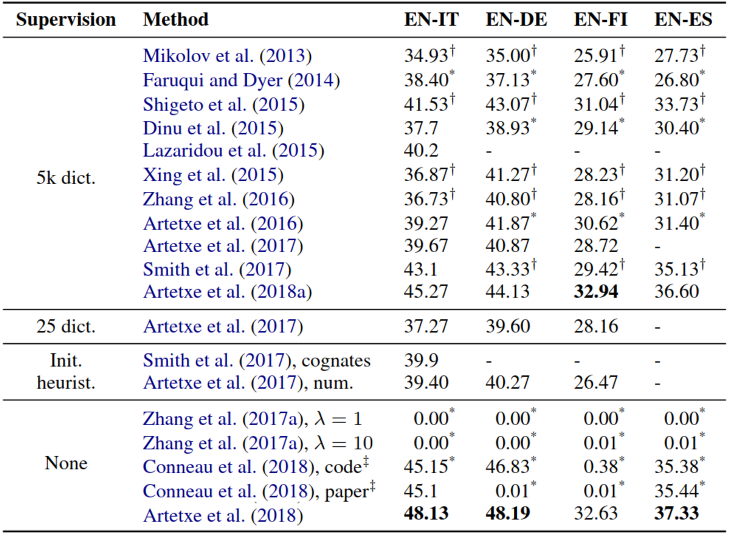
Where are we?

1. We have monolingual embeddings
2. We learn W with **adversarial training**
3. W is not good enough. Most frequent words (better embeddings) used to solve the Procrustes problem, **refined** W
4. Given a word in L1 and W, the equivalent word in L2 can be found by its nearest neighbours according to **CSLS**

Unsupervised Cross-lingual Embeddings

So, what? Comparision in Artetxe et al., ACL, 2018

- Accuracy (%)



Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Faruqui and Dyer (2014)	38.40*	37.13*	27.60*	26.80*
	Shigeto et al. (2015)	41.53†	43.07†	31.04†	33.73†
	Dinu et al. (2015)	37.7	38.93*	29.14*	30.40*
	Lazaridou et al. (2015)	40.2	-	-	-
	Xing et al. (2015)	36.87†	41.27†	28.23†	31.20†
	Zhang et al. (2016)	36.73†	40.80†	28.16†	31.07†
	Artetxe et al. (2016)	39.27	41.87*	30.62*	31.40*
	Artetxe et al. (2017)	39.67	40.87	28.72	-
	Smith et al. (2017)	43.1	43.33†	29.42†	35.13†
	Artetxe et al. (2018a)	45.27	44.13	**32.94**	36.60
25 dict.	Artetxe et al. (2017)	37.27	39.60	28.16	-
Init. heurist.	Smith et al. (2017), cognates	39.9	-	-	-
	Artetxe et al. (2017), num.	39.40	40.27	26.47	-
None	Zhang et al. (2017a), $\lambda = 1$	0.00*	0.00*	0.00*	0.00*
	Zhang et al. (2017a), $\lambda = 10$	0.00*	0.00*	0.01*	0.01*
	Conneau et al. (2018), code‡	45.15*	46.83*	0.38*	35.38*
	Conneau et al. (2018), paper‡	45.1	0.01*	0.01*	35.44*
	Artetxe et al. (2018)	**48.13**	**48.19**	32.63	**37.33**

Session II

Outline

1. Recap through the Examples of Session I

2. Word Embeddings
   - ~~Basics~~
   - ~~Frequency and Prediction-based Embeddings~~
   - Cross-lingual Embeddings

3. Unsupervised Machine Translation

Unsupervised Machine Translation

Ingredients for Today

1. Data

- Monolingual corpora

2. Initialisation

- **Cross-lingual embeddings**
- Deep MLM pretraining

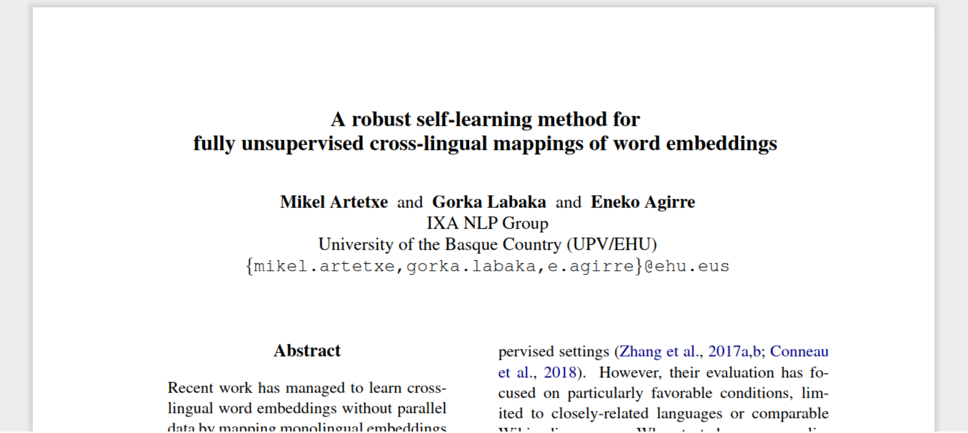
3. Training

SMT and/or **NMT**

- **Denoising autoencoder**
- **Backtranslation**

Unsupervised Machine Translation

Seminal Works by IXA (Simultaneous with Facebook)



A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings

Mikel Artetxe and **Gorka Labaka** and **Eneko Agirre**
IXA NLP Group
University of the Basque Country (UPV/EHU)
`{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus`

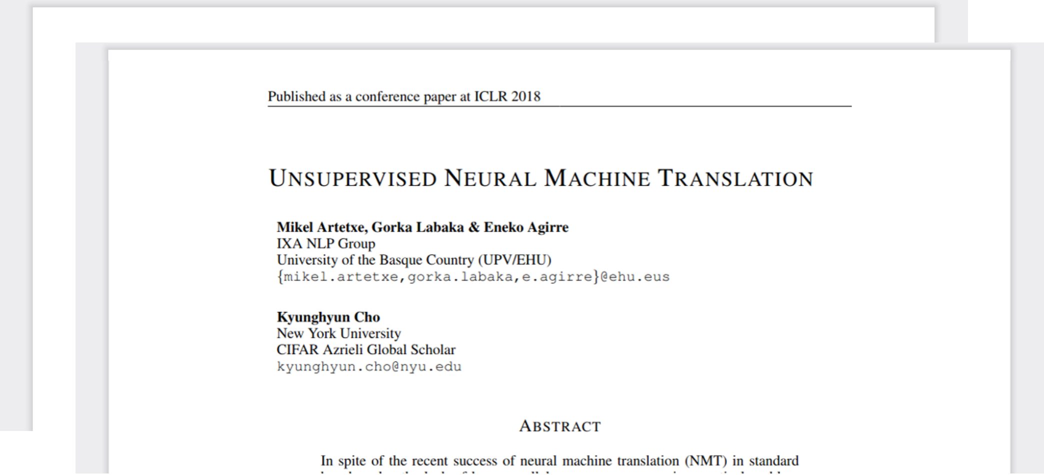
Abstract

Recent work has managed to learn cross-lingual word embeddings without parallel data by mapping monolingual embeddings

pervised settings (Zhang et al., 2017a,b; Conneau et al., 2018). However, their evaluation has focused on particularly favorable conditions, limited to closely-related languages or comparable

Unsupervised Machine Translation

Seminal Works by IXA (Simultaneous with Facebook)



Published as a conference paper at ICLR 2018

UNSUPERVISED NEURAL MACHINE TRANSLATION

Mikel Artetxe, Gorka Labaka & Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

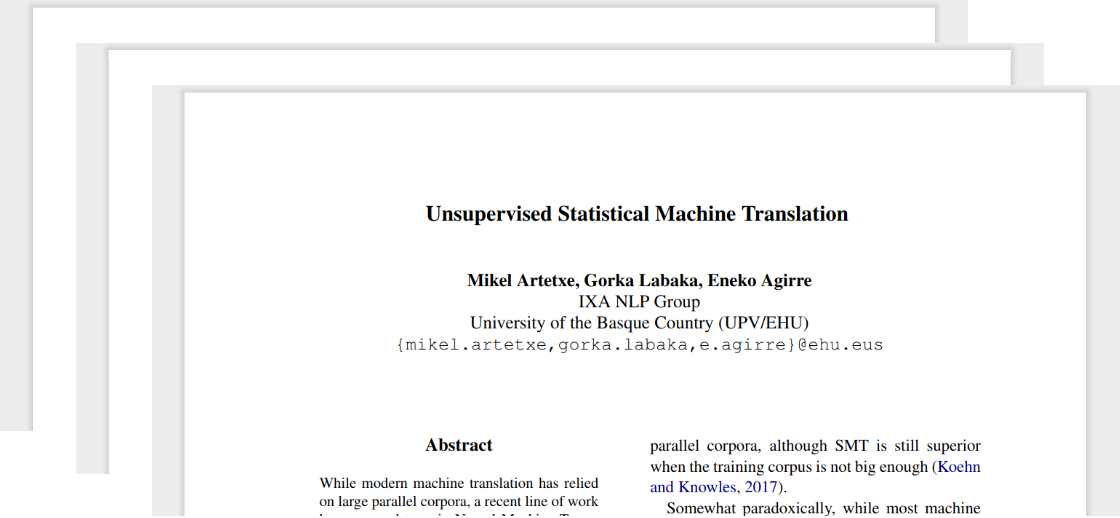
Kyunghyun Cho
New York University
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

ABSTRACT

In spite of the recent success of neural machine translation (NMT) in standard

Unsupervised Machine Translation

Seminal Works by IXA (Simultaneous with Facebook)



Unsupervised Statistical Machine Translation

Mikel Artetxe, Gorka Labaka, Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
`{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus`

Abstract

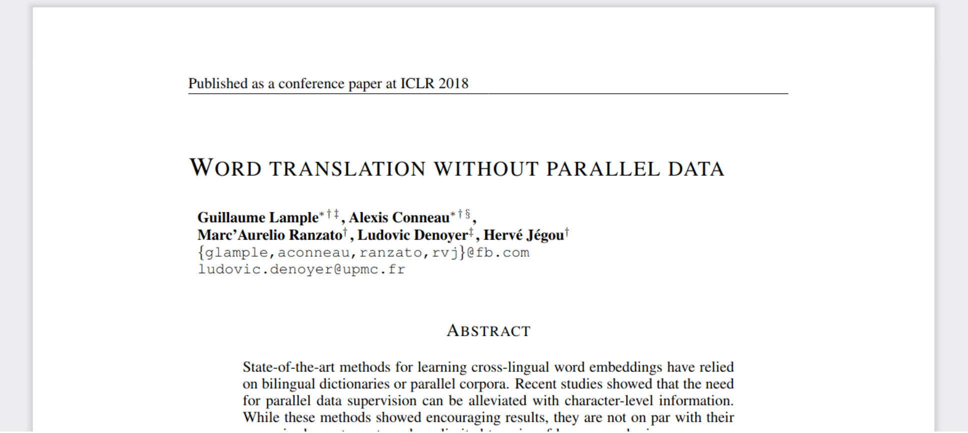
While modern machine translation has relied on large parallel corpora, a recent line of work

parallel corpora, although SMT is still superior when the training corpus is not big enough (Koehn and Knowles, 2017).

Somewhat paradoxically, while most machine

Unsupervised Machine Translation

Seminal Works by Facebook (Simultaneous with IXA)



Published as a conference paper at ICLR 2018

WORD TRANSLATION WITHOUT PARALLEL DATA

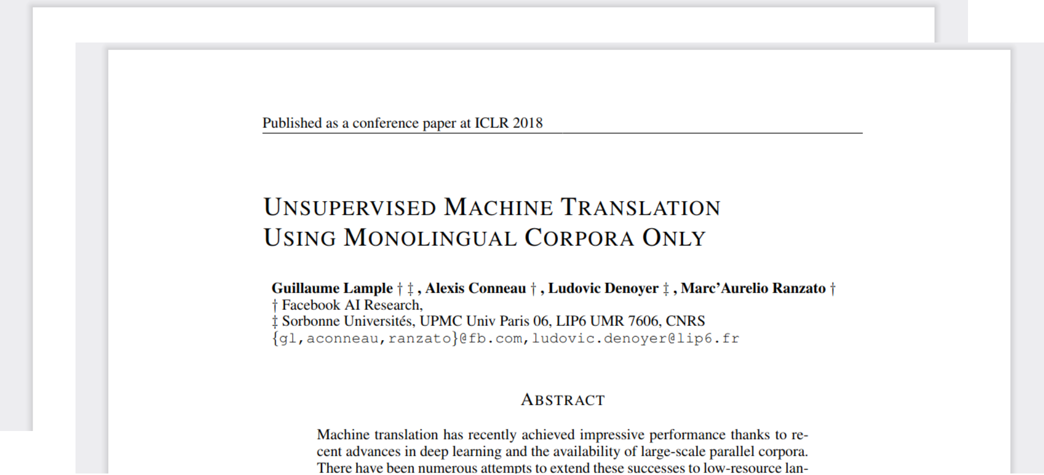
Guillaume Lample[*†‡], **Alexis Conneau**[*†§],
Marc'Aurelio Ranzato[†], **Ludovic Denoyer**[‡], **Hervé Jégou**[†]
{glample,aconneau,ranzato,rvj}@fb.com
ludovic.denoyer@upmc.fr

ABSTRACT

State-of-the-art methods for learning cross-lingual word embeddings have relied on bilingual dictionaries or parallel corpora. Recent studies showed that the need for parallel data supervision can be alleviated with character-level information. While these methods showed encouraging results, they are not on par with their

Unsupervised Machine Translation

Seminal Works by Facebook (Simultaneous with IXA)



Published as a conference paper at ICLR 2018

Unsupervised Machine Translation Using Monolingual Corpora Only

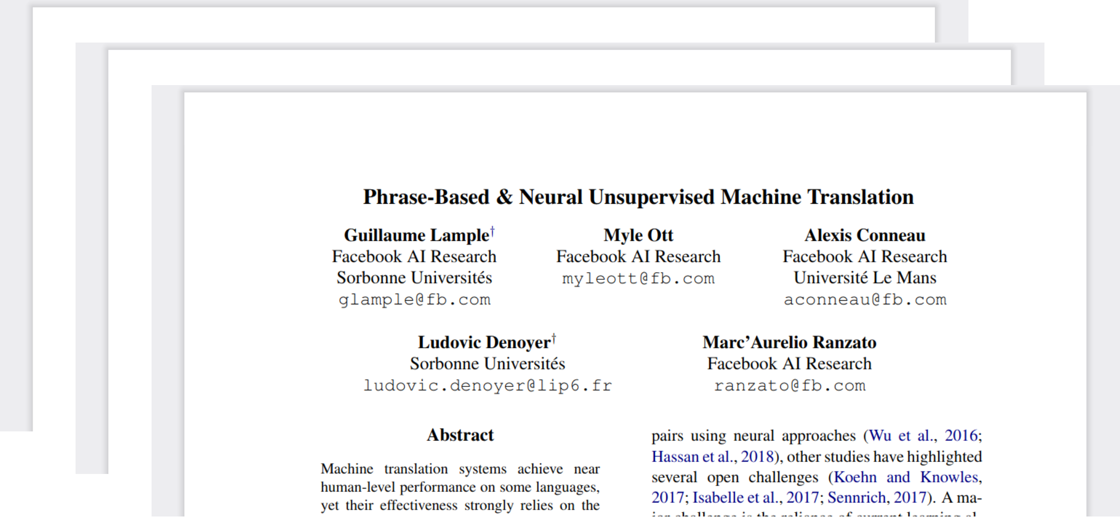
Guillaume Lample † ‡ **, Alexis Conneau** † **, Ludovic Denoyer** ‡ **, Marc'Aurelio Ranzato** †
† Facebook AI Research,
‡ Sorbonne Universités, UPMC Univ Paris 06, LIP6 UMR 7606, CNRS
`{gl,aconneau,ranzato}@fb.com,ludovic.denoyer@lip6.fr`

Abstract

Machine translation has recently achieved impressive performance thanks to recent advances in deep learning and the availability of large-scale parallel corpora. There have been numerous attempts to extend these successes to low-resource lan-

Unsupervised Machine Translation

Seminal Works by Facebook (Simultaneous with IXA)



Phrase-Based & Neural Unsupervised Machine Translation

Guillaume Lample[†]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott
Facebook AI Research
myleott@fb.com

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

Ludovic Denoyer[†]
Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc'Aurelio Ranzato
Facebook AI Research
ranzato@fb.com

Abstract

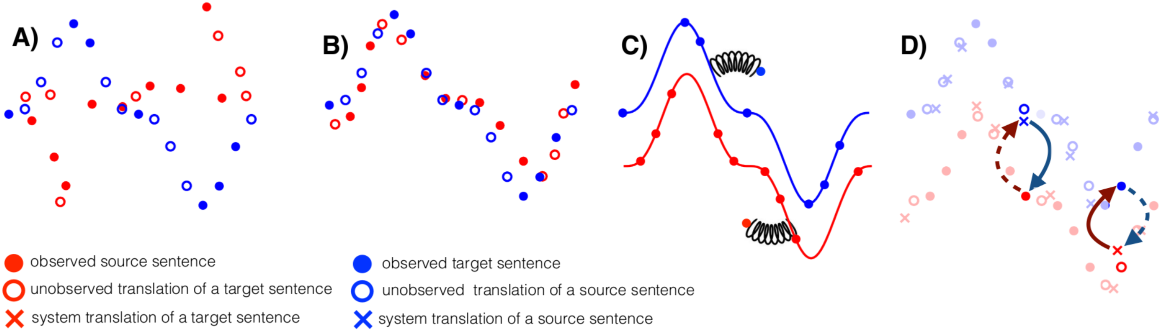
Machine translation systems achieve near human-level performance on some languages, yet their effectiveness strongly relies on the

pairs using neural approaches (Wu et al., 2016; Hassan et al., 2018), other studies have highlighted several open challenges (Koehn and Knowles, 2017; Isabelle et al., 2017; Sennrich, 2017). A major challenge is the reliance of current learning al

Unsupervised Machine Translation

The Three Principles (from Lample et al., ICLR, 2018)

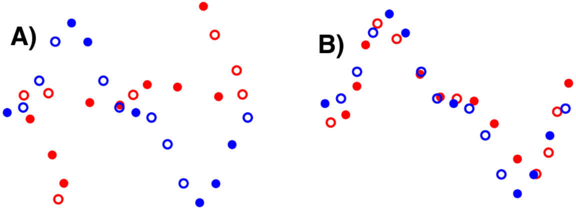
Initialisation **Denoising (LM)** **Backtranslation**

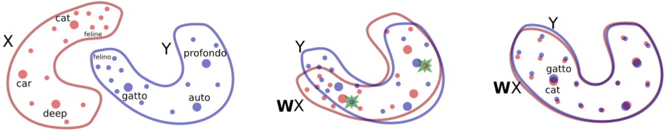

A)
B)
C)
D)
● observed source sentence
○ unobserved translation of a target sentence
✕ system translation of a target sentence
● observed target sentence
○ unobserved translation of a source sentence
✕ system translation of a source sentence

Unsupervised Machine Translation

The Three Principles (from Lample et al., ICLR, 2018)

Initialisation

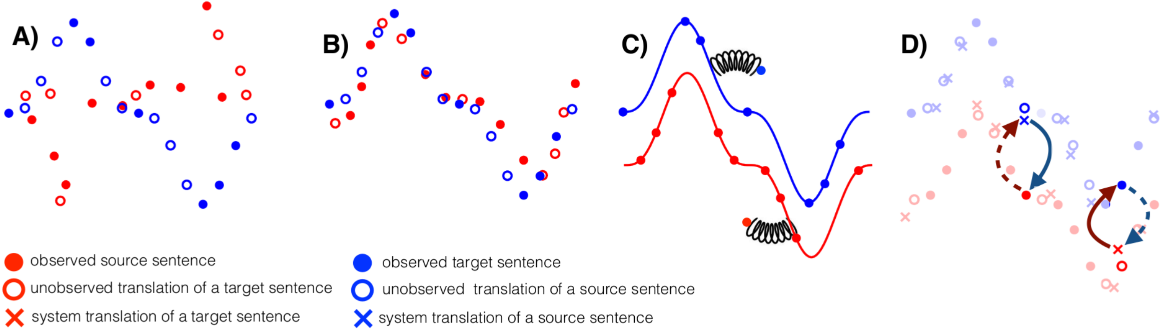




Unsupervised Machine Translation

The Three Principles (from Lample et al., ICLR, 2018)

Initialisation **Denoising (LM)** **Back-translation**


A)
B)
C)
D)
● observed source sentence
○ unobserved translation of a target sentence
× system translation of a target sentence
● observed target sentence
○ unobserved translation of a source sentence
× system translation of a source sentence

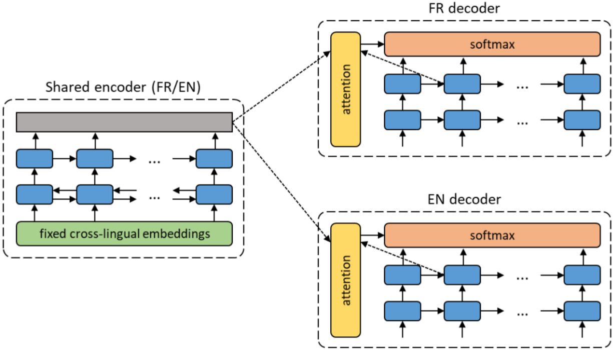
Unsupervised Machine Translation

Basics with Principles (Slides from Mikel Artetxe)

Training

- Supervised

Une fusillade a eu lieu à l'aéroport international de Los Angeles.



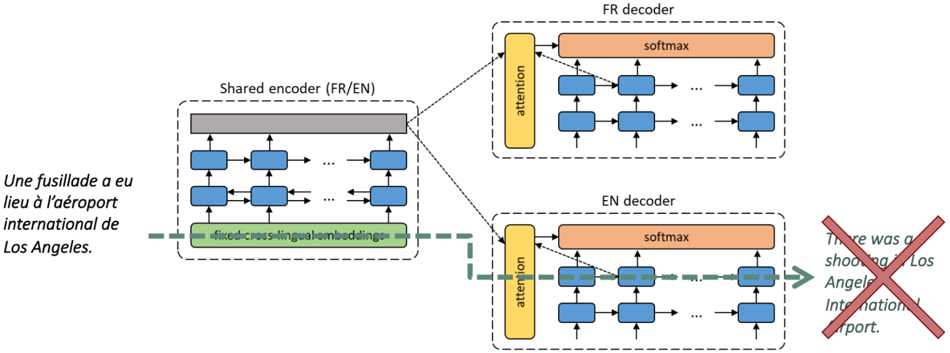
There was a shooting in Los Angeles International Airport.

Unsupervised Machine Translation

Basics with Principles (Slides from Mikel Artetxe)

Training

- Supervised



Une fusillade a eu lieu à l'aéroport international de Los Angeles.

~~There was a shooting in Los Angeles International Airport.~~

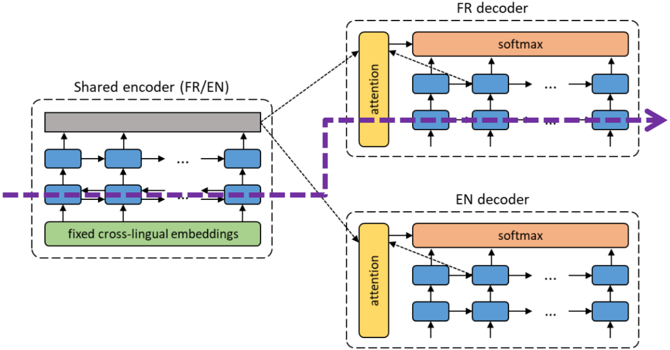
Unsupervised Machine Translation

Basics with Principles (Slides from Mikel Artetxe)

Training

- ~~Supervised~~
- Denoising

Une lieu fusillade a eu à l'aéroport de Los international Angeles.



Une fusillade a eu lieu à l'aéroport international de Los Angeles.

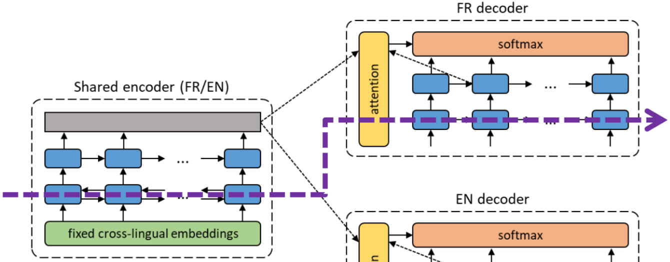
Unsupervised Machine Translation

Basics with Principles (Slides from Mikel Artetxe)

Training

- ~~Supervised~~
- Denoising

Une lieu fusillade a eu à l'aéroport de Los international Angeles.



Une fusillade a eu lieu à l'aéroport international de Los Angeles.

$$\mathcal{L}^{denoise} \sim -\log P_{s\to s}(x|C(x)) \\ -\log P_{t\to t}(y|C(y))$$

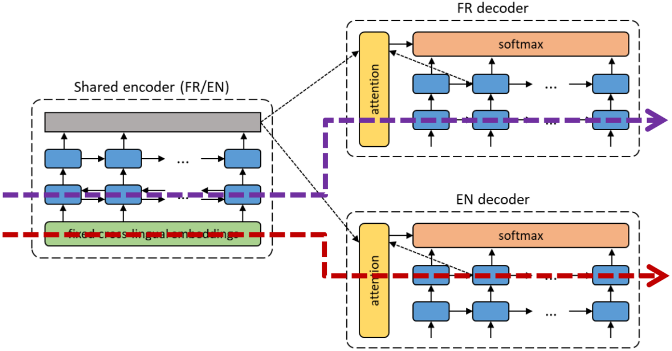
Unsupervised Machine Translation

Basics with Principles (Slides from Mikel Artetxe)

Training

- ~~Supervised~~
- Denoising
- Backtranslation

Une lieu fusillade a eu à l'aéroport de Los international Angeles.



Une fusillade a eu lieu à l'aéroport international de Los Angeles.

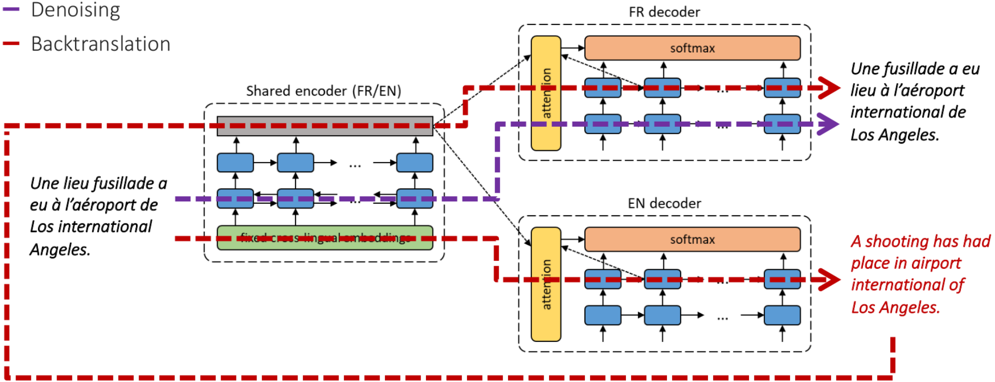
A shooting has had place in airport international of Los Angeles.

Unsupervised Machine Translation

Basics with Principles (Slides from Mikel Artetxe)

Training

- ~~Supervised~~
- Denoising
- Backtranslation



Une lieu fusillade a eu à l'aéroport de Los international Angeles.

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

A shooting has had place in airport international of Los Angeles.

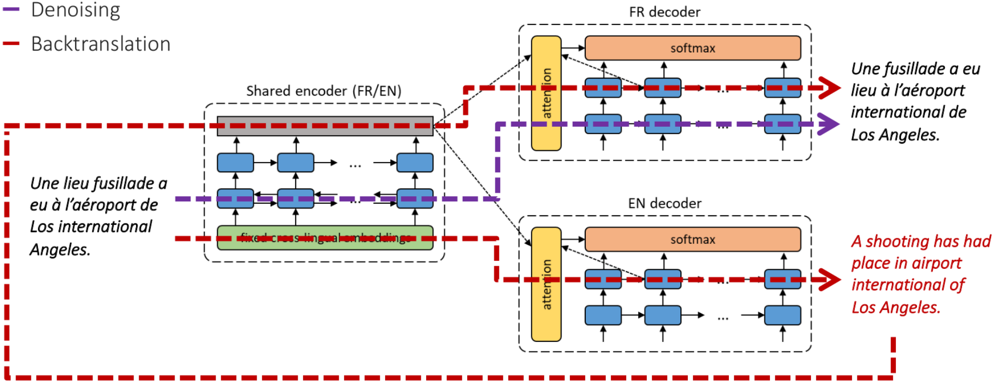
Unsupervised Machine Translation

Basics with Principles (Slides from Mikel Artetxe)

$$\mathcal{L}^{back} \sim -\log P_{s\to t}(y|u^*(y)) \\ -\log P_{t\to s}(x|v^*(x))$$

Training

- ~~Supervised~~
- Denoising
- Backtranslation



Une lieu fusillade a eu à l'aéroport de Los international Angeles.

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

A shooting has had place in airport international of Los Angeles.

Unsupervised Machine Translation

Evaluation with BLEU

		WMT-14					WMT-16	
		fr-en	en-fr	de-en	en-de		de-en	en-de
Supervised	*Vaswani et al. (2017)*	-	41.0	-	28.4	-	-	-
	Edunov et al. (2018)	-	45.6	-	35.0	-	-	-
NMT	*Artetxe et al. (2018)*	15.6	15.1	10.2	6.6		-	-
	Lample et al. (2018a)	14.3	15.1	-	-		13.3	9.6
	Lample et al. (2018b)	24.2	25.1	-	-		21.0	17.2

Unsupervised Machine Translation

Evaluation with BLEU

		WMT-14					WMT-16	
		fr-en	en-fr	de-en	en-de		de-en	en-de
Supervised	*Vaswani et al. (2017)*	-	41.0	-	28.4	-	-	-
	Edunov et al. (2018)	-	45.6	-	35.0	-	-	-
NMT	*Artetxe et al. (2018)*	15.6	15.1	10.2	6.6		-	-
	Lample et al. (2018a)	14.3	15.1	-	-		13.3	9.6
	Lample et al. (2018b)	24.2	25.1	-	-		21.0	17.2
SMT	*Artetxe et al. (2018)*	25.9	26.2	17.4	14.1		23.1	18.2
	Lample et al. (2018b)	27.2	28.1	-	-		22.9	17.9
	Artetxe et al. (2019)	28.4	30.1	20.1	15.8		25.4	19.7
SMT+ NMT	*Lample et al. (2018b)*	27.7	27.6	-	-		25.2	20.2
	Artetxe et al. (2019)	**33.5**	**36.2**	**27.0**	**22.5**		**34.4**	**26.9**

Unsupervised Machine Translation

Evaluation with BLEU

		WMT-14					WMT-16	
		fr-en	en-fr	de-en	en-de		de-en	en-de
Supervised	*Vaswani et al. (2017)*	-	41.0	-	28.4	-	-	-
	Edunov et al. (2018)	-	45.6	-	35.0	-	-	-
NMT	*Artetxe et al. (2018)*	15.6	15.1	10.2	6.6		-	-
	Lample et al. (2018a)	14.3	15.1	-	-		13.3	9.6
	Lample et al. (2018b)	24.2	25.1	-	-		21.0	17.2
SMT	*Artetxe et al. (2018)*	25.9	26.2	17.4	14.1		23.1	18.2
	Lample et al. (2018b)	27.2	28.1	-	-		22.9	17.9
	Artetxe et al. (2019)	28.4	30.1	20.1	15.8		25.4	19.7
SMT+ NMT	*Lample et al. (2018b)*	27.7	27.6	-	-		25.2	20.2
	Artetxe et al. (2019)	**33.5**	**36.2**	**27.0**	**22.5**		**34.4**	**26.9**
Leaderboard	*Unsupervised*	GPT-3	MASS	GPT-3	GPT-3		Artetxe19	Artetxe19

Unsupervised Machine Translation

An Approach for Low-Resource MT?

- No need for parallel data, only monolingual, **but**
- News Crawl 2007–2013: 749 million tokens in *fr*, 1606 in *de*, 2109 in *en*

Unsupervised Machine Translation

An Approach for Low-Resource MT?

- No need for parallel data, only monolingual, **but**
- News Crawl 2007–2013: 749 million tokens in *fr*, 1606 in *de*, 2109 in *en*

When Does Unsupervised Machine Translation Work?

Kelly Marchisio, Kevin Duhand and Philipp Koehn, WMT 2020

- on different scripts and between dissimilar languages?
- with imperfect domain alignment between source and target corpora?
- with a domain mismatch between training data and the test set?
- on the low-quality data of real low-resource languages?

Unsupervised Machine Translation

When Does Unsupervised Machine Translation Work? Marchisio et al. (2020)

	Supervised	**Parallel**	**Disjoint**	**Diff. Dom.**
Corpus	A / A	A / A	A / B	A / CC*
Ru-En	26.9	23.7 *(-3.2)*	21.2 *(-5.7)*	0.7 *(-26.2)*
Fr-En	29.9	27.6 *(-2.3)*	27.0 *(-2.9)*	3.9 *(-26.0)*

- A, B disjoint parts of UN corpus, CC (Common Crawl)
- SacreBLEU on newstest2019 (Ru-En) and newstest2014 (Fr-En)
- Different domain even more crucial than distant languages
- Why?

Unsupervised Machine Translation

When Does Unsupervised Machine Translation Work? Marchisio et al. (2020)

	Condition	**Min**	**Max**	μ	σ
Fr-En	Parallel	48.00	50.20	49.09	0.69
	Disjoint	37.88	39.09	38.47	0.37
	Diff. Dom.	**0.00**	**17.27**	**7.97**	**7.95**
	News	25.86	28.10	26.97	0.56
	CC	25.87	27.60	26.90	0.51
Ru-En	Parallel	32.24	34.04	32.95	0.47
	Disjoint	25.08	26.96	25.79	0.58
	Diff. Dom.	0.00	0.10	0.01	0.03
	News	22.19	23.77	23.10	0.44
	CC	**0.00**	**24.69**	**12.61**	**11.45**

- Accuracies (%) of induced dictionaries on 10-11 runs. Bold experiments were unstable

Unsupervised Machine Translation

When Does Unsupervised Machine Translation NOT Work? Ruiter et al. (2021)

	English	**Afrikaans**	**Nepali**	**Kannada**	**Yorúbà**	**Swahili**	**Burmese**
Typology	fusional	fusional	fusional	agglutinative	analytic	agglutinative	analytic
Word Order	SVO	SOV,SVO	SOV	SOV	SOV,SVO	SVO	SOV
Script	Latin	Latin	Brahmic	Brahmic	Latin	Latin	Brahmic
sim(L–en)	1.000	0.822	0.605	0.602	0.599	0.456	0.419

- We have see different domains (src vs. tgt, train vs. test). But also...
- When the word order is very different, different typology, different script
- All this makes mapping word embeddings a challenge

Unsupervised Machine Translation

When Does Unsupervised Machine Translation NOT Work? Ruiter et al. (2021)

Pair	Init.	Config.	Best	UMT	USMT+NMT	LASER	TSS	#P (k)
en2af	WE	B+BT	51.2±.9	27.9±.8	**44.2±.9**	52.1±1.0	35.3	37
af2en	WE	B+BT	52.2±.9	1.4±.1	**0.7±.1**	52.9±.9	–	
en2kn	DAE	B+BT+WT+N	0.3±.0	0.0±.0	**0.0±.0**	–	21.3	397
kn2en	DAE	B+BT+WT+N	0.9±.1	0.0±.0	**0.0±.0**	–	40.3	397
en2my	DAE	B(+BT+WT)	0.1±.0	0.1±.0	**0.0±.0**	0.0±.0	39.3	223
my2en	DAE	B(+BT+WT)	0.7±.1	0.0±.0	**0.0±.0**	0.1±.0	38.6	223
en2ne	DAE	B+BT+WT+N	0.3±.0	0.1±.0	**0.0±.0**	0.5±.1	8.8	–
ne2en	DAE	B+BT+WT(+N)	0.5±.0	0.0±.0	**0.0±.0**	0.2±.0	21.5	–
en2sw	WE	B+BT+WT+N	7.7±.3	3.6±.2	**0.2±.0**	10.0±.3	14.8	995
sw2en	DAE	B+BT	6.8±.2	0.3±.0	**0.0±.0**	8.4±.3	19.7	995
en2yo	WE	B+BT+WT	2.9±.1	1.0±.1	**0.3±.1**	–	12.3	501
yo2en	DAE	B+BT+WT	3.1±.1	0.6±.0	**0.0±.0**	–	22.4	501

Unsupervised Machine Translation

It's Late...

More to come!!

Thanks! And...

wait!

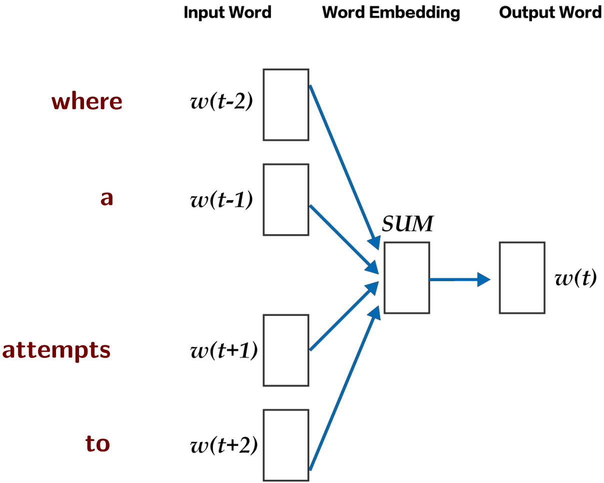
Cross-Lingual Word Embeddings
Unsupervised Machine Translation

Cristina España-Bonet
DFKI GmbH

Low-Resource NLP:
Multilinguality and Machine Translation
Webinar Series — Session II
29th June 2021

Back-up Slides

Continuous Bag of Words, CBoW

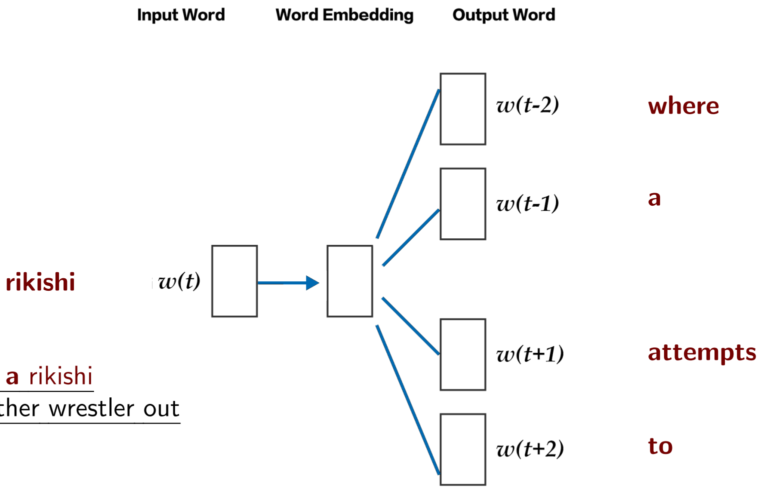


rikishi

Sumo is a sport where a **rikishi** attempts to force another wrestler out of a circular ring.

Back-up Slides

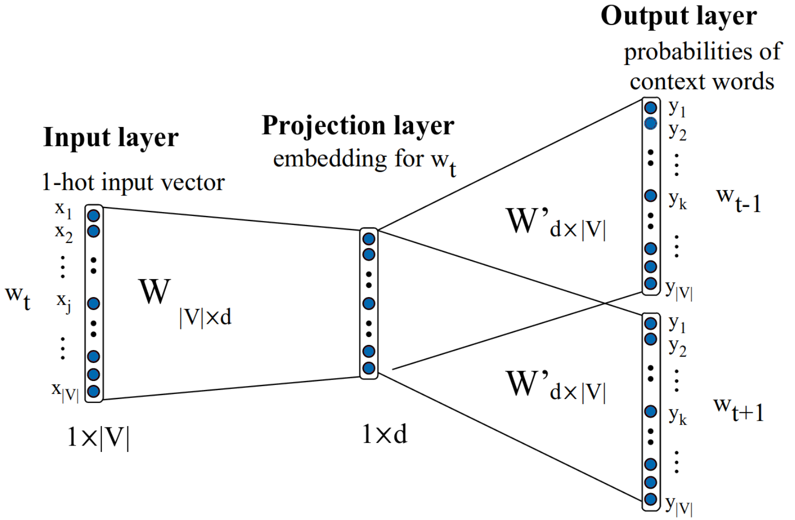
Skip-Gram Model



Sumo is a sport **where a** rikishi **attempts to** force another wrestler out of a circular ring.

Back-up Slides

More Detailed Architecture (skip-gram)



Credits: Xin Rong

Back-up Slides

More Detailed Architecture (schematic matrix visualisation)

$$\underset{\mathbf{x}}{\begin{pmatrix} \\ V \\ \\ \end{pmatrix}} \underset{\mathbf{W}}{\begin{pmatrix} \\ \quad V \times d \quad \\ \\ \end{pmatrix}} \underset{\mathbf{h}}{\begin{pmatrix} d \end{pmatrix}} \underset{\mathbf{W}'}{\begin{pmatrix} \quad d \times V \quad \end{pmatrix}} \underset{\mathbf{y}}{\begin{pmatrix} \\ V \\ \\ \end{pmatrix}}$$

Input Embedding

The row *i* of the input matrix W is the $1 \times \mathrm{d}$ for word *i* in the vocabulary

Back-up Slides

More Detailed Architecture (schematic matrix visualisation)

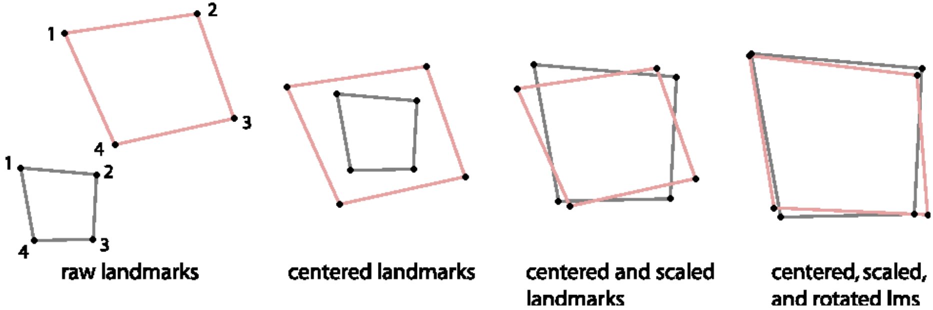
$$\underset{\mathbf{x}}{\begin{pmatrix} \\ V \\ \\ \end{pmatrix}} \underset{\mathbf{W}}{\begin{pmatrix} & & \\ & V \times d & \\ & & \end{pmatrix}} \underset{\mathbf{h}}{\begin{pmatrix} d \end{pmatrix}} \underset{\mathbf{W'}}{\begin{pmatrix} & d \times V & \end{pmatrix}} \underset{\mathbf{y}}{\begin{pmatrix} \\ V \\ \\ \end{pmatrix}}$$

Output Embedding

The column *j* of the output matrix W' is the $\mathrm{d} \times 1$ for word *j* in the vocabulary

Back-up Slides

Orthogonal Procrustes Problem


1
2
3
4
1
2
3
4
raw landmarks
centered landmarks
centered and scaled landmarks
centered, scaled, and rotated lms

Philipp Mitteroecker & Philipp Gunz, Advances in Geometric Morphometrics

Back-up Slides

Orthogonal Procrustes Problem

The proof is so simple and elegant...

$$\|AW - B\|_F = \sum_{i,j} (AW - B)^2_{i,j} =$$

$$\sum_{i,j} (AW)^2_{i,j} + (B)^2_{i,j} - 2(AW)_{i,j}(B)_{i,j} =$$

Back-up Slides

Orthogonal Procrustes Problem

The proof is so simple and elegant...

$$\|AW - B\|_F = \sum_{i,j} (AW - B)^2_{i,j} =$$

$$\sum_{i,j} (AW)^2_{i,j} + (B)^2_{i,j} - 2(AW)_{i,j}(B)_{i,j} =$$

$$\|AW\|_F + \|B\|_F - 2tr(W^T A^T B) = \|A\|_F + \|B\|_F - 2tr(W^T A^T B)$$

Back-up Slides

Orthogonal Procrustes Problem

The proof is so simple and elegant...

$$\|AW - B\|_F = \sum_{i,j} (AW - B)^2_{i,j} =$$

$$\sum_{i,j} (AW)^2_{i,j} + (B)^2_{i,j} - 2(AW)_{i,j}(B)_{i,j} =$$

$$\|AW\|_F + \|B\|_F - 2tr(W^T A^T B) = \|A\|_F + \|B\|_F - 2tr(W^T A^T B)$$

$$tr(W^T A^T B) = tr(W^T U \Sigma V^T) = (V^T W^T U \Sigma)$$

$$V^T W^T U = I \Rightarrow W = UV^T, \textbf{QED}.$$

Back-up Slides

Singular-Value Decomposition, SVD

- Linear algebra

Back-up Slides

Singular-Value Decomposition, SVD

- Linear algebra
- **Factorisation** of a matrix $\mathbf{M}$ as $\mathbf{M} = \mathbf{U\Sigma V^T}$

Back-up Slides

Singular-Value Decomposition, SVD

- Linear algebra
- **Factorisation** of a matrix $\mathbf{M}$ as $\mathbf{M} = \mathbf{U\Sigma V^T}$
    - *D* $\mathbf{U}$ is an $m \times m$ orthogonal matrix,

Back-up Slides

Singular-Value Decomposition, SVD

- Linear algebra
- **Factorisation** of a matrix $\mathbf{M}$ as $\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}$

  D $\mathbf{U}$ is an $m \times m$ **orthogonal matrix**,

  - $\mathbf{U}^{\mathbf{T}}\mathbf{U} = \mathbf{U}\mathbf{U}^{\mathbf{T}} = \mathbf{I}$
  - or, equivalently, $\mathbf{U}^{\mathbf{T}} = \mathbf{U}^{-1}$

Back-up Slides

Singular-Value Decomposition, SVD

- Linear algebra
- **Factorisation** of a matrix $\mathbf{M}$ as $\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}}$
  - *D* $\mathbf{U}$ is an $m \times m$ orthogonal matrix,
  - *D* $\boldsymbol{\Sigma}$ is a diagonal $m \times n$ matrix with non-negative real numbers,

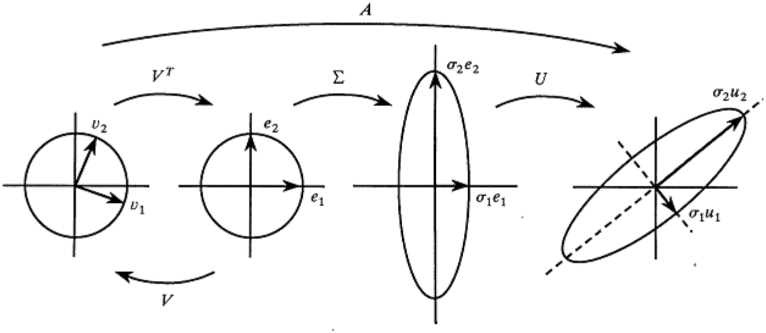
Back-up Slides

Singular-Value Decomposition, SVD

- Linear algebra
- **Factorisation** of a matrix $\mathbf{M}$ as $\mathbf{M} = \mathbf{U\Sigma V^T}$
    - *D* $\mathbf{U}$ is an $m \times m$ orthogonal matrix,
    - *D* $\mathbf{\Sigma}$ is a diagonal $m \times n$ matrix with non-negative real numbers,
    - *D* $\mathbf{V^T}$ is the conjugate transpose of an $n \times n$ orthogonal matrix

Back-up Slides

SVD: 2×2 Geometric Interpretation



a linear transformation is a rotation or reflection, followed by a scaling, followed by another rotation or reflection

`https://blogs.sas.com`

Back-up Slides

Singular-Value Decomposition, SVD

$$\begin{pmatrix} \\ \\ m \times n \\ \\ \\ \end{pmatrix} = \begin{pmatrix} \\ \\ \quad\quad m \times m \quad\quad \\ \\ \\ \end{pmatrix} \begin{pmatrix} \\ \\ m \times n \\ \\ \\ \end{pmatrix} \begin{pmatrix} \\ n \times n \\ \\ \end{pmatrix}$$

$$\mathbf{M} \quad = \quad \mathbf{U} \quad\quad \boldsymbol{\Sigma} \quad\quad \mathbf{V}^{\mathbf{T}}$$

Back-up Slides

SVD: Singular Values

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & & & \\ & \ddots & \text{\Large 0} & \\ & \text{\Large 0} & \sigma_r & \\ & & & 0 \end{pmatrix};$$

σ_1 ... σ_r, singular values of **M** (in decreasing order)

r, rank of **M**

Back-up Slides

SVD: Singular Values

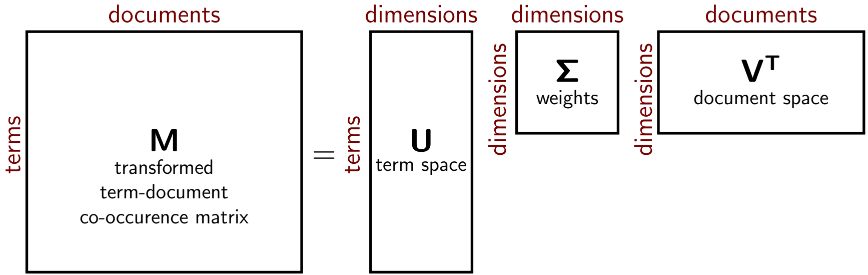
$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & & & & \\ & . & & 0 & \\ & & . & & \\ & 0 & & \sigma_r & \\ & & & & 0 \end{pmatrix} ; \qquad \mathbf{M_r} = \sum_{i=1}^{r} \sigma_i \overrightarrow{u}_i \overrightarrow{v}_i^T$$

σ_1 ... σ_r, singular values of **M** (in decreasing order)

r, rank of **M**

Back-up Slides

SVD: Application, Latent Semantic Analysis



documents

terms

M

transformed

term-document

co-occurence matrix

=

dimensions

terms

U

term space

dimensions

dimensions

Σ

weights

documents

dimensions

$\mathbf{V}^\mathbf{T}$

document space

$$\Sigma_r \Longrightarrow \mathbf{M}_r$$

Back-up Slides

SVD: Learn & Practice

`https://nlp.stanford.edu/IR-book/pdf/18lsi.pdf`

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. **Introduction to Information Retrieval**. Cambridge University Press, New York, NY, USA.