Deep Learning for NLP

Universität Bielefeld

Lecture 6 – Word Embeddings 2 (Syntactic, Bilingual, Contextualized Embeddings)

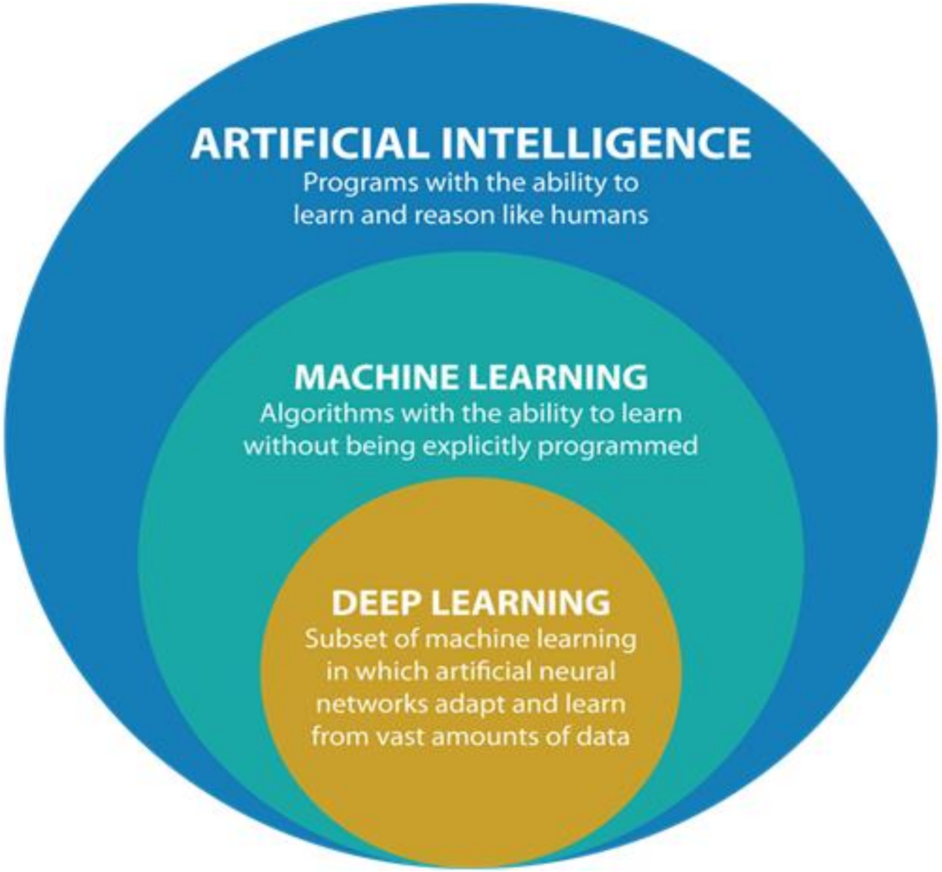
Dr. Steffen Eger
steffen.eger@uni-bielefeld.de

Natural Language Learning Group (NLLG)

CITEC NLLG

Recap

- Started out with general ML
- Then saw DL as a subfield of ML



Recap

Universität Bielefeld

- We then talked about NLP
- Today, a large part of NLP is
  - Learning **tasks** from (human) **labeled datasets**
  - Input is **text**

Recap

Universität Bielefeld

- We then talked about NLP
- Today, a large part of NLP is
  - Learning **tasks** from (human) **labeled datasets**

Recap

Universität Bielefeld

- We then talked about NLP
- Today, a large part of NLP is
  - Learning **tasks** from (human) **labeled datasets**

Text	Label
Buy Viagra at 5$	Spam
I like soccer	No-Spam
All the world‘s a stage	No-Spam
…	…

Recap

Universität Bielefeld

- We then talked about NLP
- Today, a large part of NLP is
  - Learning **tasks** from (human) **labeled datasets**

Text	**Label**
Where there is a "will," there are 500 relatives	Funny
I like ice-cream	Not funny
Always remember: you're unique, just like everyone else	Funny
…	…

NLLG

Recap

Universität Bielefeld

- We then talked about NLP
- Today, a large part of NLP is
  - Learning **tasks** from (human) **labeled datasets**

Text1	Text2	Label
I like cats	I like dogs	Similarity: high
I like ice-cream	Bielefeld is a city	Similarity: low
Dallas Mavericks will win	Zverev lost again	Similarity: medium
…		…

NLLG

Recap

Universität Bielefeld

- We then talked about NLP
- Today, a large part of NLP is
  - Learning **tasks** from (human) **labeled datasets**

Text1	Text2	Label
I like cats	I like dogs	Adequacy: 0.2
I like ice-cream	I enjoy to eat my ice-cream	Adequacy: 0.7
Dallas Mavericks will win	Zverev lost again	Adequacy: 0.1
…		…

NLLG

Recap

Universität Bielefeld

- We then talked about NLP
- Today, a large part of NLP is
  - Learning **tasks** from (human) **labeled datasets**

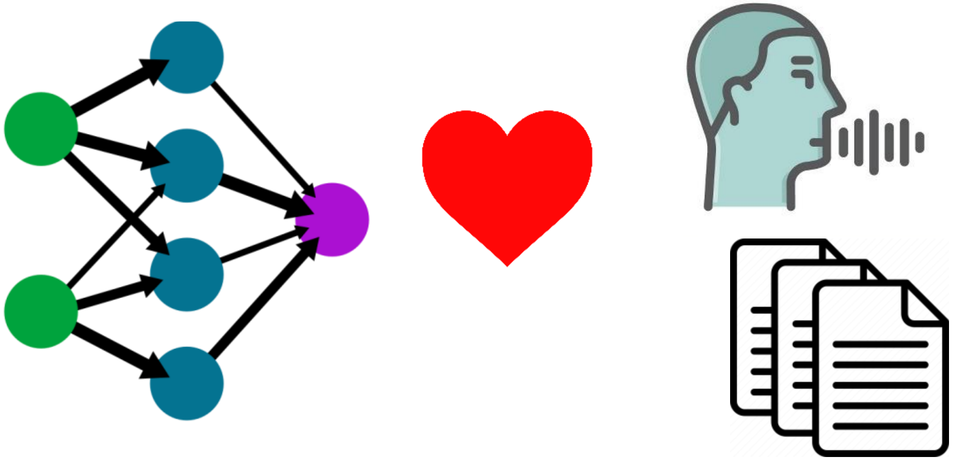
Text	Label
I like cats	PRON VERB NOUN
I like ice-cream	PRON VERB NOUN
Dallas Mavericks will win	Name Name VERB VERB
…	…

NLLG

Recap

- Combining both worlds
- We need vector representations for text inputs

  → **representation learning**

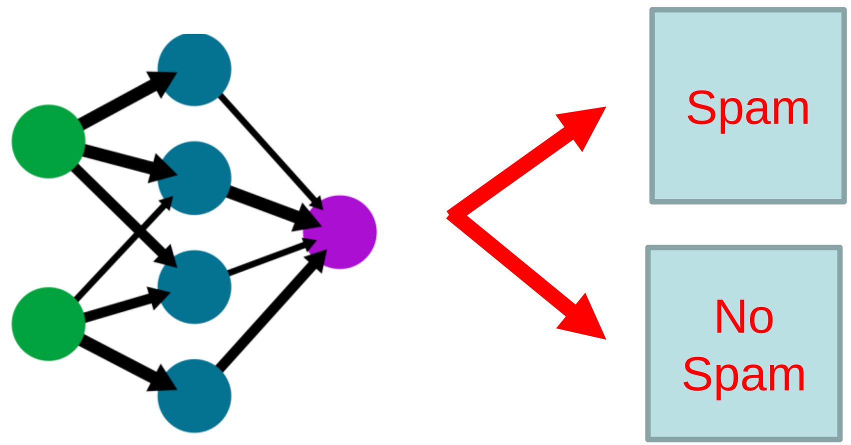


Recap

- Combining both worlds
- We need vector representations for text inputs

  → **representation learning**

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$



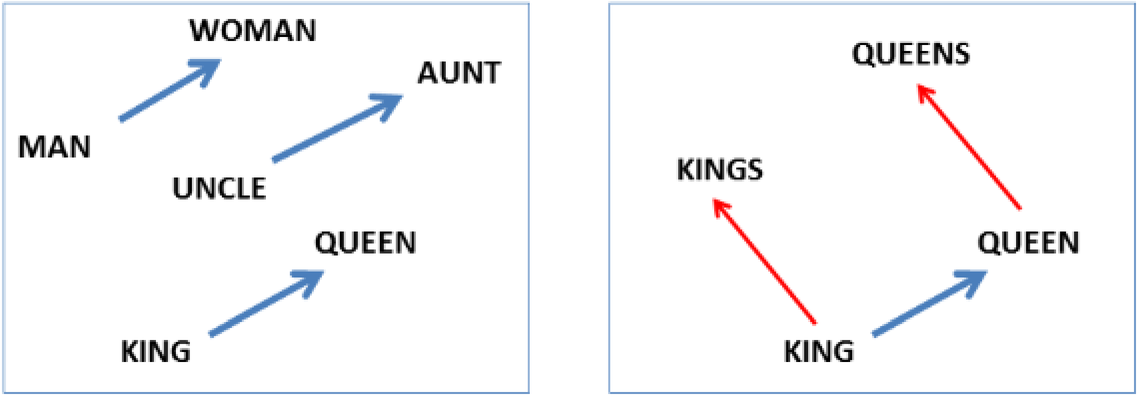
Quiz

The recap was …

A: .. waste of time. I knew this already
B: .. an eye-opener. Finally I know what‘s going on
C: .. helpful for sure! (Even though I knew the big picture already)
D: .. useless! I‘m still lost!

Last session

- Word embeddings can represent semantic and syntactic relations between words in the vector space



Mikolov et al (2013a)

Linguistic Regularities in Continuous Space Word Representations

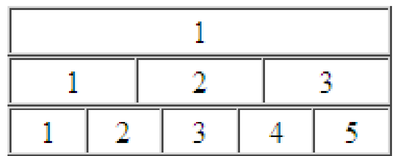
This lecture

1) **Multi-Sense Embeddings**
2) Multi-Lingual Embeddings
3) Syntactic Word Embeddings
4) Other aspects
5) Contextualized Embeddings

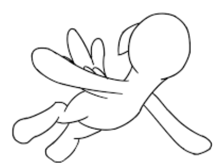
Word Senses

Universität Bielefeld

- Words do not represent only one meaning



…



You are pretty fly…

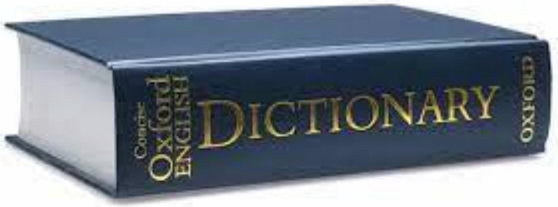
- Problem is generally known as *polysemy* (or even *homonymy*): a word may have many different meanings

Word Senses

Man
1. The human species
2. Males of the human species
3. Adult males of the human species

Bank
1. A financial institution
2. The building where a financial institution offers services
3. A synonym for „rely upon“
4. **Note:** River *bank* is a homonym to 1 and 2

book
1. A bound collection of pages
2. A text reproduced and distributed
3. Make an action or event a matter of record



Sense-disambiguated word representations

- Idea: Train word vectors on sense-disambiguated corpora

  Example from the [SemCor] corpus:

```
<s snum=132>
<wf cmd=ignore pos=DT>A</wf>
<wf cmd=done pos=NN lemma=rush wnsn=2 lexsn=1:11:00::>rush</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done pos=NN lemma=panic wnsn=1 lexsn=1:12:00::>panic</wf>
<wf cmd=done pos=VB lemma=catch wnsn=12 lexsn=2:30:00::>caught</wf>
<wf cmd=done rdf=person pos=NNP lemma=person wnsn=1 lexsn=1:03:00::
pn=person>Sarah</wf>
<punc>.</punc>
</s>
```

Sense-disambiguated word representations

- Idea: Train word vectors on sense-disambiguated corpora

  Example from the SemCor corpus:

  <s snum=132>

  <wf cmd=ignore pos=DT>**A**</wf>

  <wf cmd=done pos=NN lemma=rush wnsn=2 lexsn=1:11:00::>**rush**</wf>

  <wf cmd=ignore pos=IN>**of**</wf>

  <wf cmd=done pos=NN lemma=panic wnsn=1 lexsn=1:12:00::>**panic**</wf>

  <wf cmd=done pos=VB lemma=catch wnsn=12 lexsn=2:30:00::>**caught**</wf>

  <wf cmd=done rdf=person pos=NNP lemma=person wnsn=1 lexsn=1:03:00:: pn=person>**Sarah**</wf>

  <punc>.</punc>

  </s>

  A rush_2 of panic_1 caught_12 Sarah_1

Sense-disambiguated word representations

Universität Bielefeld

- Result: different representations for each sense

$\textit{bank}^n_1$ (geographical)	$\textit{bank}^n_2$ (financial)	$\textit{number}^n_4$ (phone)	$\textit{number}^n_3$ (acting)	$\textit{hood}^n_1$ (gang)	$\textit{hood}^n_{12}$ (convertible car)
upstream^r_1	$\text{commercial\_bank}^n_1$	calls^n_1	appearing^v_6	tortures^n_5	taillights^n_1
downstream^r_1	$\text{financial\_institution}^n_1$	dialled^v_1	minor_roles^n_1	vengeance^n_1	grille^n_2
runs^v_6	national_bank^n_1	operator^n_{20}	$\text{stage\_production}^n_1$	badguy^n_1	bumper^n_2
confluence^n_1	trust_company^n_1	$\text{telephone\_network}^n_1$	$\text{supporting\_roles}^n_1$	brutal^a_1	fascia^n_2
river^n_1	savings_bank^n_1	telephony^n_1	leading_roles^n_1	execution^n_1	rear_window^n_1
stream^n_1	banking^n_1	subscriber^n_2	stage_shows^n_1	murders^n_1	headlights^n_1

Table 1: Closest senses to two senses of three ambiguous nouns: *bank*, *number*, and *hood*

- Iacobacci et al (2015): *SensEmbed: Learning Sense Embeddings for Word and Relational Similarity*

NLLG

DISCUSS

Universität Bielefeld

How to train an NLP system with these sense-disambiguated embeddings?

NLLG

A more parsimonious approach

- Run word2vec on data and compute embeddings
- For each target word, represent its context as avg. or concatenated embedding

... need to go to the bank to get some money ….

… debt by utilizing a credit line granted by a bank …

…. raw water is largely river bank filtrate (approximately 70 percent) …

… runs from its idyllic river bank promenade under the Elbe to …

A more parsimonious approach

Universität Bielefeld

- Run word2vec on data and compute embeddings
- For each target word, represent its context as avg. or concatenated embedding

... need to go to the bank to get some money ….

… debt by utilizing a credit line granted by a bank …

…. raw water is largely river bank filtrate (approximately 70 percent) …

… runs from its idyllic river bank promenade under the Elbe to …

A more parsimonious approach

... need to go to the bank to get some money ….

context = [.2,.8]

... debt by utilizing a credit line granted by a bank …

context = [.4,.6]

…. raw water is largely river bank filtrate (approximately 70 percent) …

context = [-.2,-.8]

… runs from its idyllic river bank promenade under the Elbe to …

context = [-.9,-.3]

- Cluster the context representations, and assign each word's context to a cluster → the word has the sense corresponding to the cluster index
  - Using techniques from *unsupervised* machine learning (see lecture 2)
- Run word2vec on sense-disambiguated corpus

A more parsimonious approach

... need to go to the bank to get some money

... debt by u

.... raw wat

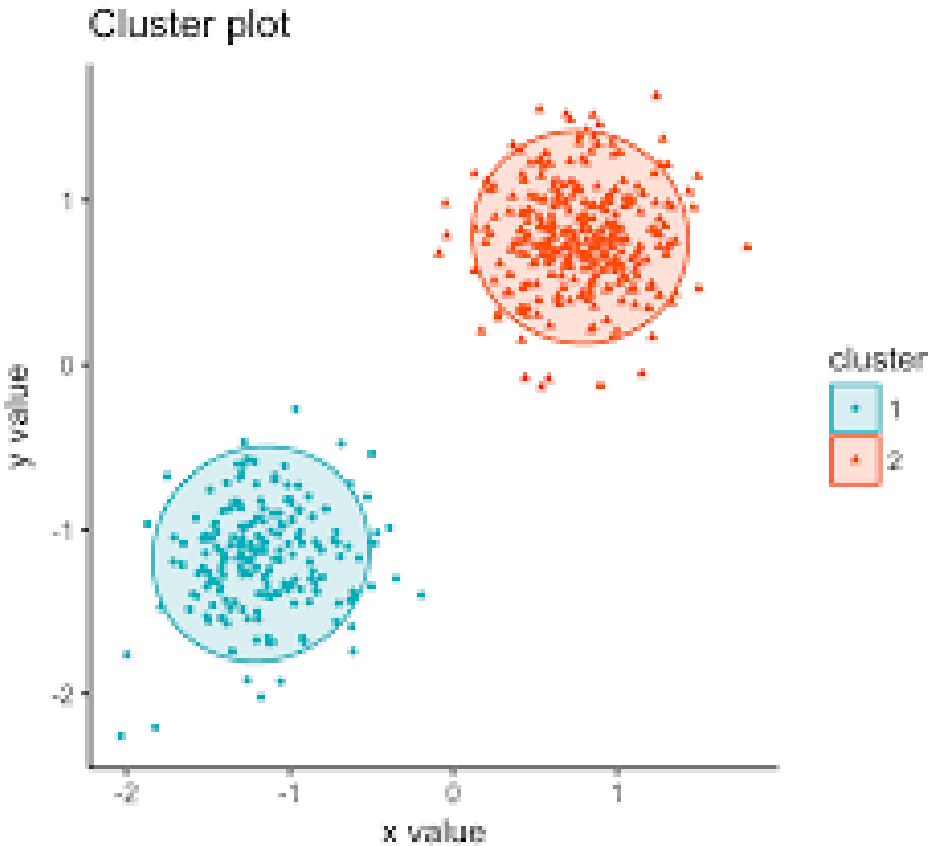
... runs fron

context = [.2,.8]

context = [.4,.6]

ely 70 percent) ... context = [-.2,-.8]

e Elbe to ... context = [-.9,-.3]



- Cluster the co ... s context to a cluster → the ... ster index
  - Using tec ... (see lecture 2)
- Run word2vec

Sense-disambiguated word representations

Universität Bielefeld

However, in practice, most people didn’t use sense embeddings

- Not so much benefit in using them in practical applications
- On the other hand, the cost is much higher --- one needs a sense-labeler or a computation heavy model

- Before ELMo and BERT came around in 2018 (see below) …
  - With **contextualized word embeddings**

NLLG

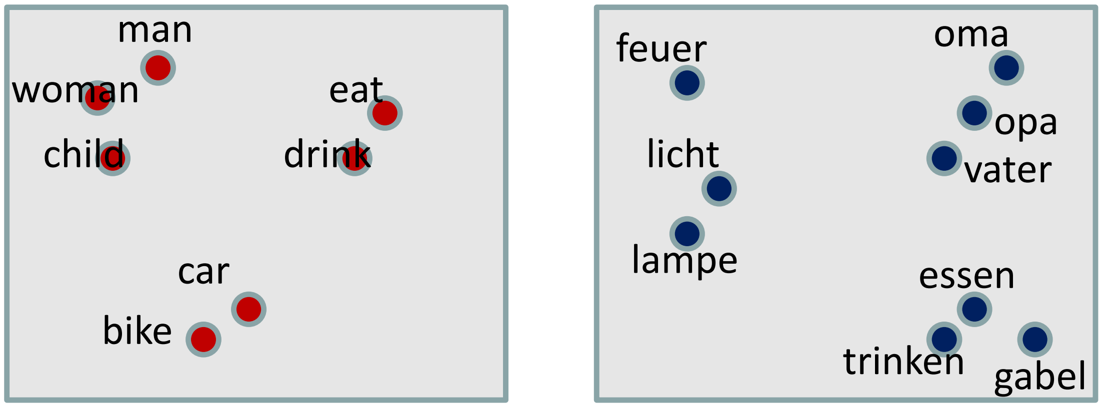
This lecture

1) Multi-Sense Embeddings
2) **Multi-Lingual Embeddings**
3) Syntactic Word Embeddings
4) Other aspects
5) Contextualized embeddings

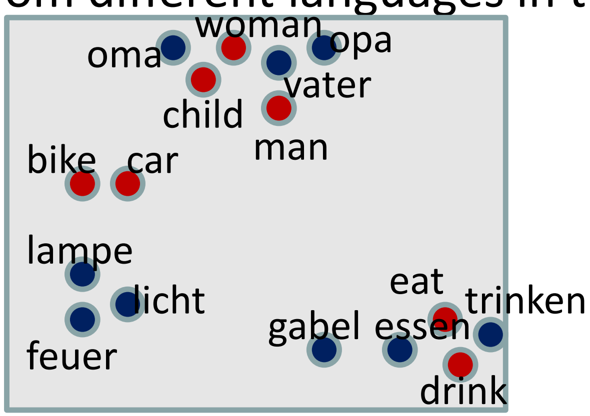
Bilingual Embeddings

- Word representations for two languages:

  → train on corpus from each language



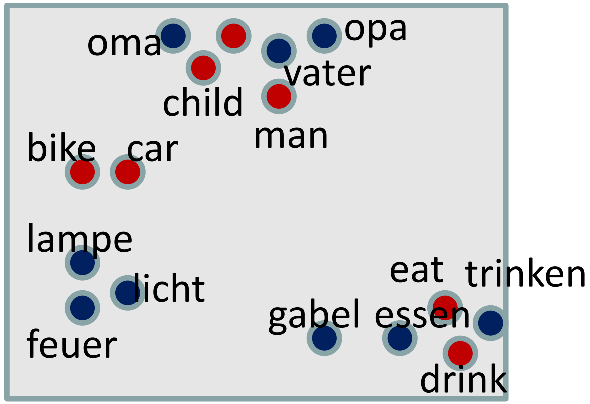
- Goal: represent words from different languages in the same space



Bilingual Embeddings

Universität Bielefeld

Goal: represent words from different languages in the same space



Bilingual Embeddings – General idea

Universität Bielefeld

- Can think of it as having two objectives we want to satisfy

- **cross-lingual objective**: words that are translations of each other should be close in the projected space
- **mono-lingual objective**: words that occur in monolingually similar contexts should be close to each other in vector space

NLLG

Bilinguality – Why?

Universität Bielefeld

(1) Second language may act as an additional “signal”

- Which may help to improve word embeddings even in the first language
  - → **Make Monolingual Embeddings better**

- E.g. assume that some word like “opa” occurs very infrequently in the German corpus, thus it’s difficult to reliably estimate its word embedding
- If its English translation “grandfather” occurs frequently in the English corpus, the German word should get a more appropriate embedding in the bilingual space

Bilinguality – Why?

(2) If words are projected in a common space (“shared features”), this may allow for **Direct Transfer / Zero-shot learning**

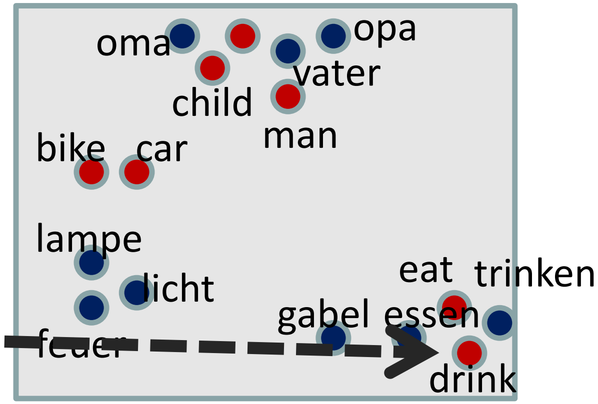
- Train a model in one language (usually resource-rich)
- Directly apply in another language (usually resource-poor)

Bilinguality – Example

Universität Bielefeld

(2) Example Direct Transfer: task is POS tagging

- Goal / approach:
  - *Train:* I may not drink this → PRON VERB PARTICLE VERB DET
  - *Test*: Es ist wichtig, ausreichend zu trinken → ….
- Training (idea):
  - Input: center words with their context words
  - Output: labels of center word
  - E.g. (not,drink,this) → VERB



- **Direct transfer / zero-shot learning:**
  - train using bilingual embeddings in English (assume big labeled English dataset)
  - then apply to German data

NLLG

DISCUSS

Universität Bielefeld

Name problems of the zero-shot learning approach. When and why will it not perform well?

Approach 1: Learning a transformation matrix

Universität Bielefeld

- One of the first and simplest approaches

  Mikolov et al. 2013, Exploiting similarities among languages for machine translation

- Given: monolingual embeddings + dictionary
  - Dictionary: *cat-Katze*, *table-Tisch*, …

x_i	z_i
cat	Katze
table	Tisch
…	…

NLLG

Approach 1: Learning a transformation matrix

- One of the first and simplest approaches

  Mikolov et al. 2013, Exploiting similarities among languages for machine translation

- Given: monolingual embeddings + dictionary
  - Dictionary: *cat-Katze*, *table-Tisch*, …

x_i	z_i
[0.2,-0.3,0.8]	[0.5,0.9,-1]
[1,2,-5]	[0.1,-0.1,0.1]
…	…

Approach 1: Learning a transformation matrix

Universität Bielefeld

- We estimate a linear transformation from this data:

$$\min_{\boldsymbol{W}} \sum_i ||\boldsymbol{x}_i \boldsymbol{W} - \boldsymbol{z}_i||^2$$

  - $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are monolingual vectors of words from dictionary

- Once $\boldsymbol{W}$ is learned, we can map any language x word into the space of language z
  - Even words for which we do not have translations

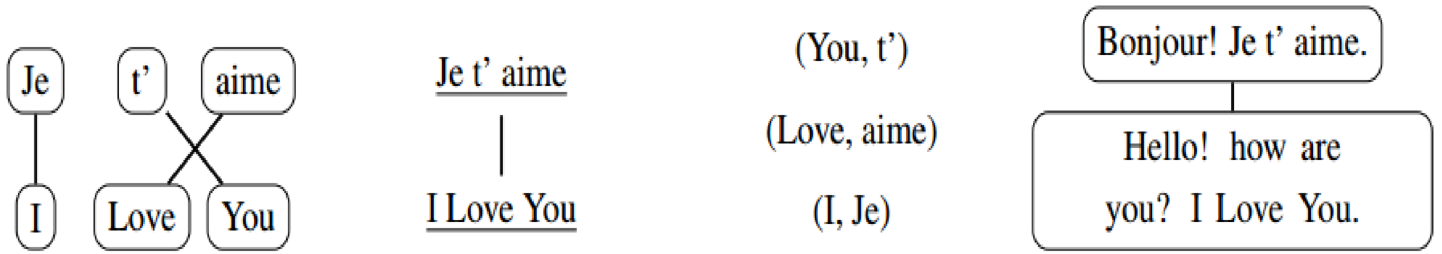
NLLG

More Bilingual Embeddings

Universität Bielefeld

- See Upadhayay et al. (2016)
  - Cross-lingual Models of Word Embeddings: An Empirical Comparison
- And more recent Glavas et al. (2019)
  - How to (properly) evaluate cross-lingual word embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions

More Bilingual Embeddings



(a) **BiSkip** (b) **BiCVM** (c) **BiCCA** (d) **BiVCD**

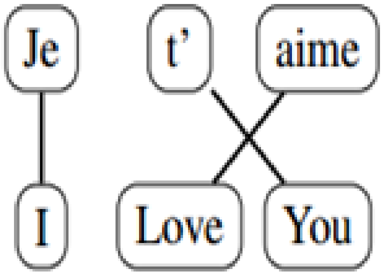
Figure 2: Forms of supervision required by the four models compared in this paper. From left to right, the cost of the supervision required varies from expensive (BiSkip) to cheap (BiVCD). BiSkip requires a parallel corpus annotated with word alignments (Fig. 2a), BiCVM requires a sentence-aligned corpus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

Bilingual Embeddings

Universität Bielefeld

- We discuss (a) BiSkip and (d) BiVCD

- **BiSkip** uses sentence and word aligned texts, then runs a skip-gram model whose contexts are words from both languages:
  - E.g. on input *love* BiSkip wants to predict the context *je, I*, *you, t'*;
  - similar for *aime: t', you*
  - → similar contexts are predicted → similar representations



(a) **BiSkip**

Determining alignments (for BiSkip)

Universität Bielefeld

- Word/Sentence alignments learned from parallel corpora

Determining bi-lingual mappings

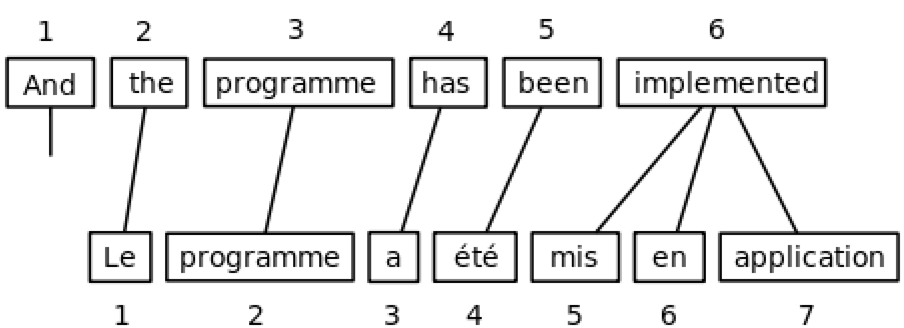
- Word/Sentence alignments learned from parallel corpora

Europarl: parallel corpus from the European parliament visualized by IMS:

Also , I would like to pay tribute to the clarity of the report and to the innovations it suggests , which are the result of **deep** analysis .

Daher möchte ich die Klarheit des vorliegenden Berichts und seine Neuerungsvorschläge hervorheben , die die Frucht intensiver Überlegungen sind .	Aussi voudrais -je rendre hommage à la clarté du rapport présenté et aux innovations qu' il propose et qui sont le résultat d' une réflexion en profondeur .	Por eso quisiera rendir homenaje a la claridad del informe presentado y a las innovaciones que propone y que son el resultado de una reflexión a fondo .	Vorrei anche rendere omaggio alla chiarezza della relazione presentata e alle innovazioni che propone , e che sono il risultato di una riflessione approfondita .

Learn word alignments

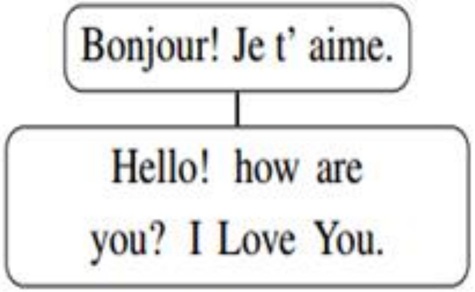


Bilingual Embeddings

Universität Bielefeld

- We discuss (a) BiSkip and (d) BiVCD

- **BiVCD** is even simpler. Given aligned documents (e.g. Wikipedia articles)
  - Merge them, then random shuffle all words
  - Then run a Monolingual Model (e.g. CBOW, Glove, Skip-Gram) on it
  - Why does this yield meaningful results?



(d) **BiVCD**

NLLG

Multilinguality

Universität Bielefeld

- We talked about mapping two languages in a common space
- How about 3, 5, 10 languages?
- Early work: Ammar et al. (2016), Massively Multilingual word embeddings
  - They extend BiCCA to MultiCCA and BiSkip to MultiSkip
- In recent years, people use **Multilingual BERT** (MBERT), which yields embeddings in a joint space for 100+ languages

More recent trends

- Learn bilingual word embeddings from as few resources as possible,
  - E.g., dictionary with only 10 word pairs (can be punctuation)
- E.g. Artexte et al., Learning bilingual word embeddings with (almost) no bilingual data, ACL 2017
  - From there we can go to unsupervised machine translation

More recent trends

- E.g. Artexte et al., Learning bilingual word embeddings with (almost) no bilingual data, ACL 2017
- Main idea:
  - If we had a dictionary, we can get bilingual embeddings
  - If we had bilingual embeddings, we can get a dictionary

More recent trends

Universität Bielefeld

- E.g. Artexte et al., Learning bilingual word embeddings with (almost) no bilingual data, ACL 2017
- Idea:
  - 1) Use a lexicon (*seed lexicon* is easy to get automatically)
  - 2) Learn bilingual embeddings using current lexicon (→ Mikolov’s method, i.e., “Approach 1”)
  - 3) Get a better lexicon using bilingual embeddings
  - 4) Go back to 1)

More recent trends

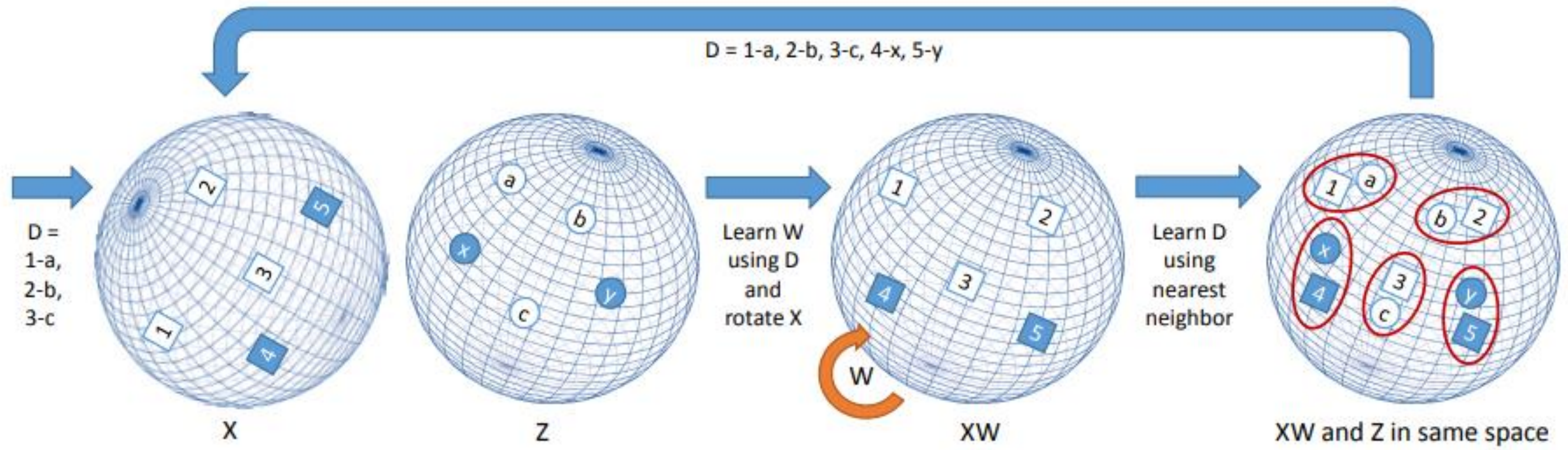


Figure 1: A general schema of the proposed self-learning framework. Previous works learn a mapping W based on the seed dictionary D, which is then used to learn the full dictionary. In our proposal we use the new dictionary to learn a new mapping, iterating until convergence.

This lecture

This lecture

1) Multi-Sense Embeddings
2) Multi-Lingual Embeddings
3) **Syntactic Word Embeddings**
4) Other aspects
5) Contextualized Embeddings

Long-distance dependencies

Universität Bielefeld

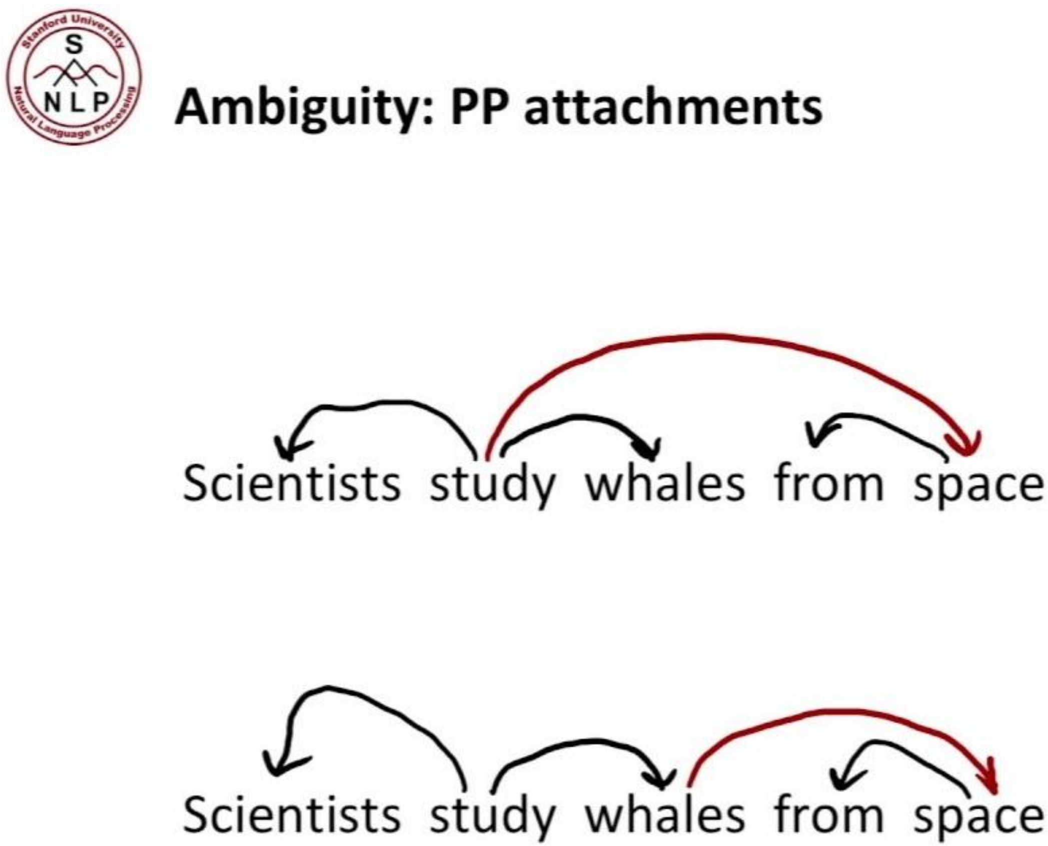
- Words can be similar with respect to (grammatical) role in a sentence
  - tea/milk/beer/coffee can all be an object of the verb *drink*

- Words that share syntactic relations might be distant in a sentence:

*I would like to **drink** a very hot tall decaf half-soy (…) white chocolate **mocha***

Dependency parsing in one slide

Outlines grammatical **relationships** between words in a sentence



Dependency parses

- Idea: apply dependency parsing first

*I would like to **drink** a very hot tall decaf half-soy (…) white chocolate **mocha***

Output of Stanford dependency parser:

nsubj(like-3, I-1)	nsubj(drink-5, I-1)	aux(like-3, would-2)
root(ROOT-0, like-3)	mark(drink-5, to-4)	xcomp(like-3, drink-5)
det(mocha-14, a-6)	advmod(hot-8, very-7)	amod(mocha-14, hot-8)
amod(mocha-14, tall-9)	amod(mocha-14, decaf-10)	amod(mocha-14, half-soy-11)
amod(mocha-14, white-12)	compound(mocha-14, chocolate-13)	
dobj(drink-5, mocha-14)		

Dependency-based embeddings

I would like to drink a very hot tall decaf half-soy (…) white chocolate mocha

nsubj(like-3, I-1)
nsubj(drink-5, I-1)
aux(like-3, would-2)
root(ROOT-0, like-3)
mark(drink-5, to-4)
xcomp(like-3, drink-5)
det(mocha-14, a-6)
advmod(hot-8, very-7)
amod(mocha-14, hot-8)
amod(mocha-14, tall-9)
amod(mocha-14, decaf-10)
amod(mocha-14, half-soy-11)
amod(mocha-14, white-12)
compound(mocha-14, chocolate-13)
dobj(drink-5, mocha-14)

- Levy and Goldberg, 2014: *Dependency-Based Word Embeddings*

Word	Dependency Context
like	I/nsubj, would/aux, drink/xcomp
drink	I/nsubj, to/mark, mocha/dobj, like/xcomp^{-1}
hot	very/advmod, mocha/amod^{-1}
…	…

Dependency-based embeddings

- Word2Vec finds words that **associate with** other words, while Dependency Embeddings finds words **behave like** others
  - *Domain similarity* vs. *functional similarity*

Target Word	BoW5	BoW2	Deps
batman	nightwing	superman	superman
	aquaman	superboy	superboy
	catwoman	aquaman	supergirl
	superman	catwoman	catwoman
	manhunter	batgirl	aquaman
hogwarts	dumbledore	evernight	sunnydale
	hallows	sunnydale	collinwood
	half-blood	garderobe	calarts
	malfoy	blandings	greendale
	snape	collinwood	millfield
turing	nondeterministic	non-deterministic	pauling
	non-deterministic	finite-state	hotelling
	computability	nondeterministic	heting
	deterministic	buchi	lessing
	finite-state	primality	hamming
florida	gainesville	fla	texas
	fla	alabama	louisiana
	jacksonville	gainesville	georgia
	tampa	tallahassee	california
	lauderdale	texas	carolina
	aspect-oriented	aspect-oriented	event-driven

Dependency-based embeddings

Universität Bielefeld

- Word2Vec finds words that **associate with** other words, while Dependency Embeddings finds words **behave like** others
  - *Domain similarity* vs. *functional similarity*

Target Word	BoW5	BoW2	Deps
batman	nightwing	superman	superman
	aquaman	superboy	superboy
	catwoman	aquaman	supergirl
	superman	catwoman	catwoman
	manhunter	batgirl	aquaman
hogwarts	dumbledore	evernight	sunnydale
	hallows	sunnydale	collinwood
	half-blood	garderobe	calarts
	malfoy	blandings	greendale
	snape	collinwood	millfield
turing	nondeterministic	non-deterministic	pauling
	non-deterministic	finite-state	hotelling
	computability	nondeterministic	heting
	deterministic	buchi	lessing
	finite-state	primality	hamming
florida	gainesville	fla	texas
	fla	alabama	louisiana
	jacksonville	gainesville	georgia
	tampa	tallahassee	california
	lauderdale	texas	carolina
	aspect-oriented	aspect-oriented	event-driven

More syntactically oriented embeddings

Universität Bielefeld

- **Syntactic** relations between words should also be represented in the vectors

  → Problem: word order matters

Dog bites man. vs Man bites dog.

Position Information

- Remember: The word2vec models do not consider position information:
  - No distinction between left and right context
  - No distinction between close and far contexts

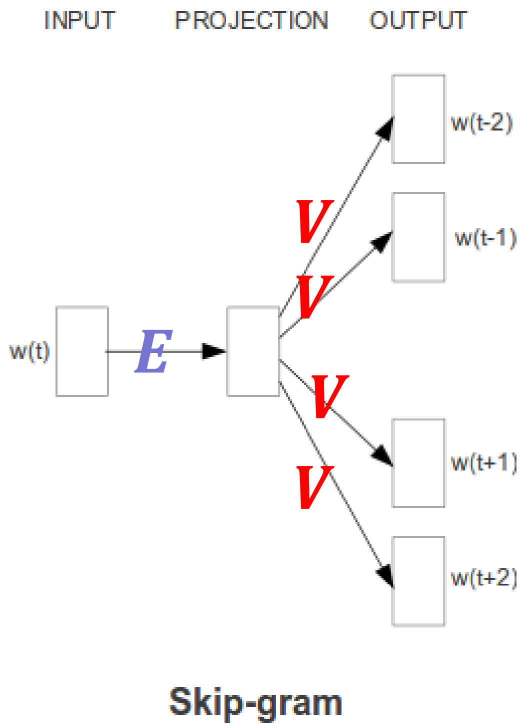
Skip-gram: ___ *bites* ___

→ *(bites, man) , (bites, dog)*

- dog bites man vs man bites dog
  - *(bites, dog-1), (bites, man+1) vs (bites, man-1), (bites, dog+1)*

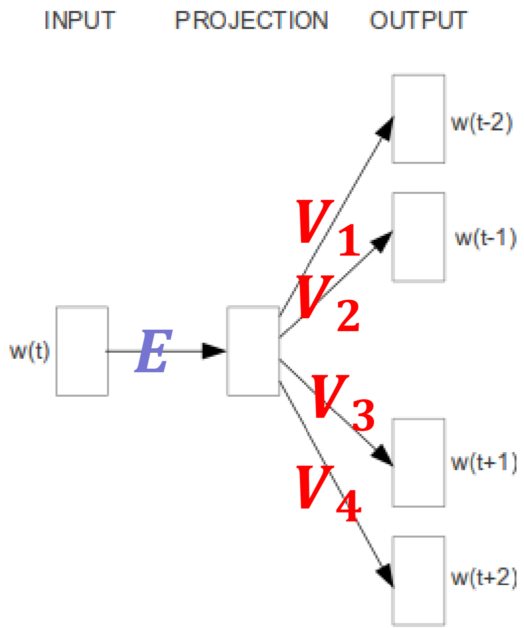
- This is "intuitively" what we want (although we don't add indices to words; why?)

The Skip-gram model



How can we predict different words when V is always the same?

The Structured Skip-gram model



Structured Skip-gram

Results

Universität Bielefeld

- Nearest neighbors for "*breaking*"

Skip-gram	Structured Skip-gram
breaks	*putting*
turning	*turning*
broke	*sticking*
break	*pulling*
stumbled	*picking*

- Word representations with positional information work slightly better for syntactic tasks like POS-tagging and parsing

- Ling et al. 2015: *Two/Too Simple Adaptations of Word2Vec for Syntax Problems*

This lecture

Universität Bielefeld

1) Multi-Sense Embeddings
2) Multi-Lingual Embeddings
3) Syntactic Word Embeddings
4) **Other aspects**
5) Contextualized Embeddings

Embeddings and Lexical Resources

Universität Bielefeld

Several NLP researchers have proposed to combine

- NLP (linguistic) resources (which e.g. capture meaning) with
- the now classical word vectors

- Faruqui et al. (2015) combine resources such a the paraphrase database (PPDB) with Embeddings
  - PPDB lists synonyms, extracted from bi-lingual datasets

$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j)\in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$

NLLG

Embeddings and Lexical Resources

Universität Bielefeld

Several NLP researchers have proposed to combine

- NLP (linguistic) resources (which e.g. capture meaning) with
- the now classical word vectors

- Faruqui et al. (2015) combine resources such a the paraphrase database (PPDB) with Embeddings
    - PPDB lists synonyms, extracted from bi-lingual datasets

Original word vector

$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j)\in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$

NLLG

Embeddings and Lexical Resources

Universität Bielefeld

Several NLP researchers have proposed to combine

- NLP (linguistic) resources (which e.g. capture meaning) with
- the now classical word vectors

- Faruqui et al. (2015) combine resources such a the paraphrase database (PPDB) with Embeddings
  - PPDB lists synonyms, extracted from bi-lingual datasets

New word vector

$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j)\in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$

NLLG

Embeddings of other things than words

Universität Bielefeld

Embed other things than words:

- **Characters**: *i n s i g h t f u l*
  - However, there are no pre-trained embeddings on the net, why?
- Or **syllables**: *in + sight + ful*
- Or **morphemes**:
  - *insightful = insight + ful*
  - *helping = help + ing*
  - *greedily = greedy + ly*
  - *Dampfschifffahrt = Dampf+Schiff+Fahrt*
  - Useful (?) particulary for morphologically rich languages like
    - German, French, Czech, etc.
    - Rarely find *Dampfschifffahrt* in a corpus, but its three morphemes are quite likely
- Embed **postags, synsets, lexemes, supersenses** (Flekova and Gurevych, 2016), …

Embeddings of other things than words

Universität Bielefeld

- Embed **n-grams**
  - That's the **FastText** approach
  - Bojanowski et al. 2016, Enriching Word Vectors with Subword Information
  - Very popular, available in many languages

- Words are represented as bags of character n-grams (n=3,4,5,6)

  E.g., n=3: where = (>wh , whe, her, ere , re<)
- Embeddings for all n-grams are learned
- Representation for a word is given by average over its n-gram embeddings

- Big advantage:
  - Can embed OOV words, e.g. spelling mistakes: "lenght", "spellling"
  - Naturally works for morphologically rich languages

This lecture

1) Multi-Sense Embeddings
2) Multi-Lingual Embeddings
3) Syntactic Word Embeddings
4) Other aspects
5) **Contextualized Embeddings**

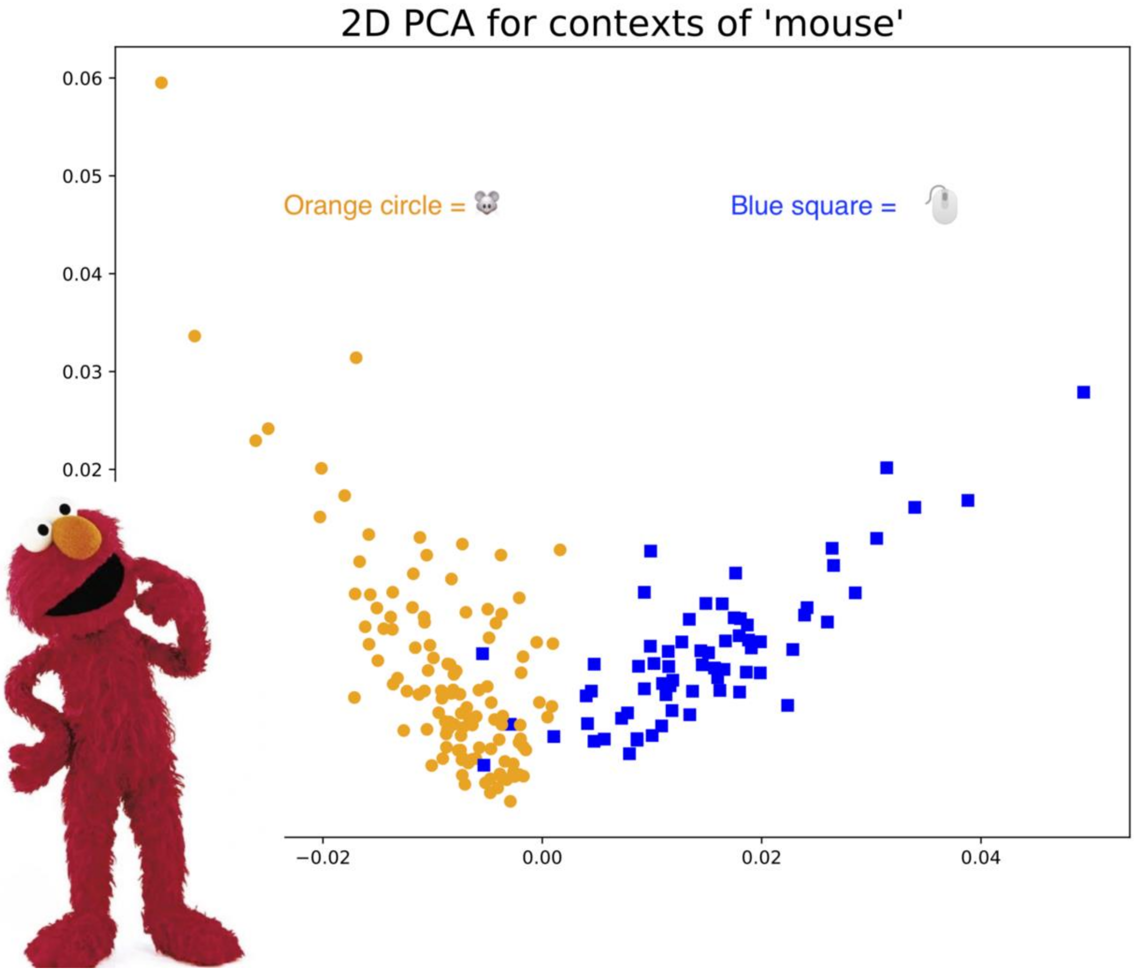
Contextualized word embeddings: ELMo & BERT

Universität Bielefeld

- ELMo and BERT use language models to get **contextualized word representations**: in each context a word has a different embedding
- They are the absolute methods of choice at the moment

NLLG

Contextualized word embeddings: ELMo & BERT



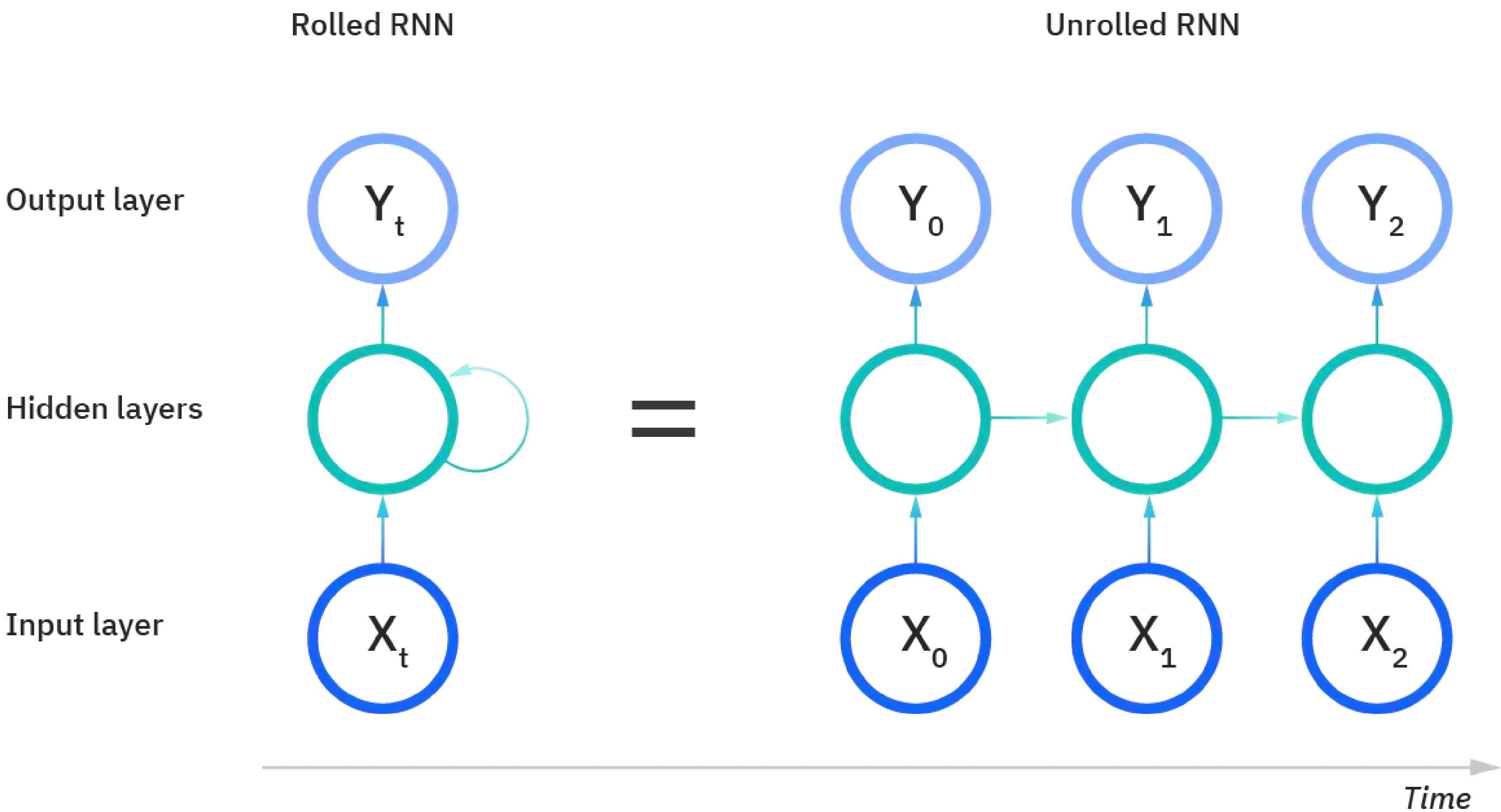
Contextualized word embeddings: ELMo & BERT

Universität Bielefeld

- ELMo combines three representations:
  - One on character level
  - Two representations obtained from the two layers in an RNN

- The language model is pre-trained on a large corpus
- For a new task, weights for the three representations are learned to get a task-specific representation
- This task specific representation is concatenated with standard static word embeddings

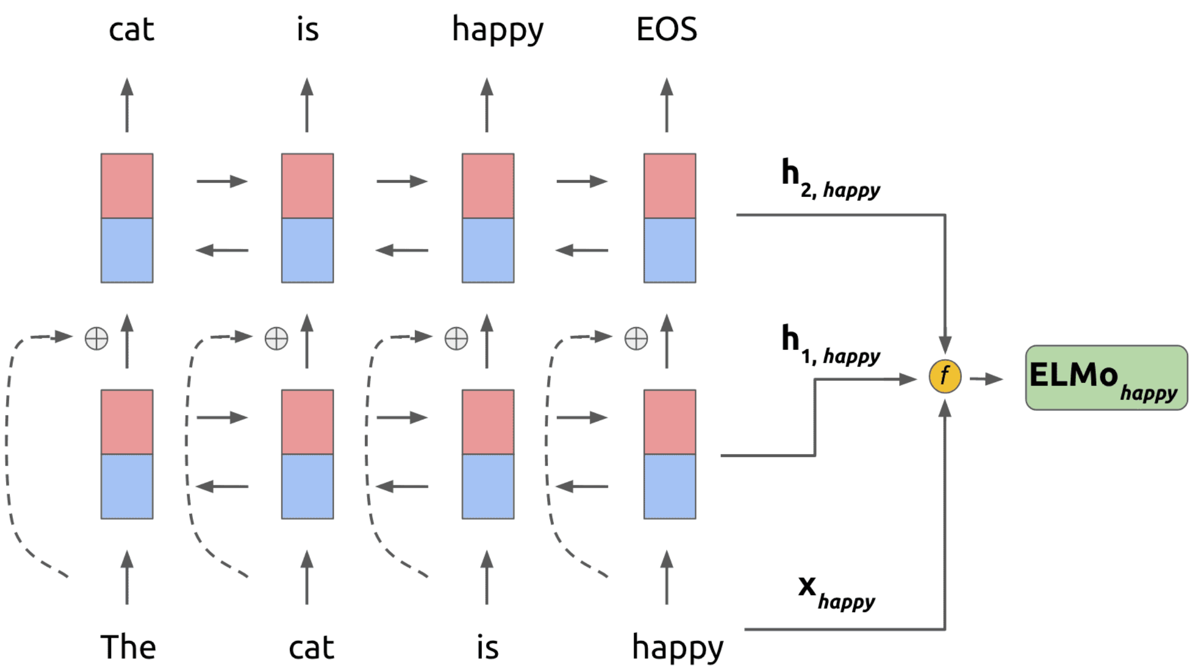
ELMo: Recurrent Neural Networks



Contextualized word embeddings: ELMo & BERT

Universität Bielefeld

- ELMo visually:



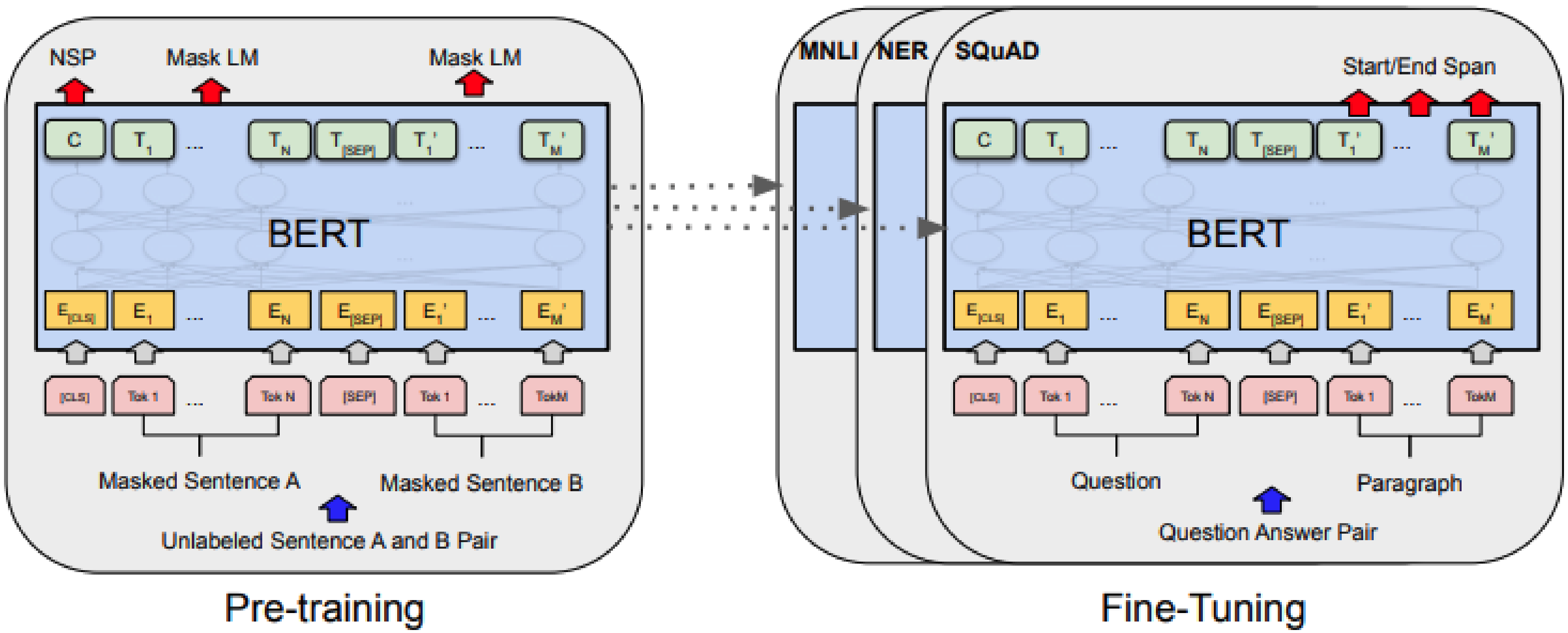
- ELMo (mid-2018) outperformed static word embeddings considerably
- More on ELMo: https://www.mihaileric.com/posts/deep-contextualized-word-representations-elmo/

NLLG

Contextualized word embeddings: BERT

Universität Bielefeld

BERT has changed NLP fundamentally: pre-training & fine-tuning



Use BERT for all kinds of tasks:

https://github.com/huggingface/pytorch-pretrained-BERT

Quiz

What are main differences between ELMo and BERT?

A: BERT uses Transformers
B: ELMo computes static representations
C: BERT combines representation learning and downstream task modeling, ELMo doesn‘t
D: BERT is deep, ELMo is shallow
E: Both use Masked Language Modeling

Summary: Embedding approaches

Universität Bielefeld

- What do all the embedding approaches have in common?
  → Represent natural language input with real-valued vectors
- Differences:
  - Unit of representation:
    - characters, morphemes, words, senses, phrases, sentences, documents, …
  - Definition of context for training:
    - CBOW, Skip-gram, Glove, positional, dependency-based, …
  - Monolinguality or multi-linguality
  - Combination with classical lexical resources

- Static vs. Contextualized representations

Summary: Embedding approaches

Universität Bielefeld

- Note that **static word embeddings are becoming extinct** now
- … and replaced by **contextualized embeddings**

References (1)

Universität Bielefeld

- Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. "SensEmbed: Learning Sense Embeddings for Word and Relational Similarity." *ACL (1)*. 2015.
- Huang, Eric H., et al. "Improving word representations via global context and multiple word prototypes." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012.
- Neelakantan, Arvind, et al. "Efficient non-parametric estimation of multiple embeddings per word in vector space." *arXiv preprint arXiv:1504.06654* (2015).
- Luong, Thang, Hieu Pham, and Christopher D. Manning. "Bilingual word representations with monolingual quality in mind." *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. 2015.
- Hermann, Karl Moritz, and Phil Blunsom. "Multilingual distributed representations without word alignment." *arXiv preprint arXiv:1312.6173* (2013).
- Vulic, Ivan, and Marie-Francine Moens. "Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction." *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*. ACL, 2015.

References (2)

- Upadhyay, Shyam, et al. "Cross-lingual models of word embeddings: An empirical comparison." *arXiv preprint arXiv:1604.00425* (2016).
- Klementiev, Alexandre, Ivan Titov, and Binod Bhattarai. "Inducing crosslingual distributed representations of words." (2012).
- Upadhyay, Shyam, et al. "Cross-lingual models of word embeddings: An empirical comparison." *arXiv preprint arXiv:1604.00425* (2016).
- Bengio, Yoshua, and Greg Corrado. "Bilbowa: Fast bilingual distributed representations without word alignments." (2015).
- Ling et al. 2015: Two/Too Simple Adaptations of Word2Vec for Syntax Problems
- Levy and Goldberg, 2014: Dependency-Based Word Embeddings
- Komninos, Alexandros, and Suresh Manandhar. "Dependency based embeddings for sentence classification tasks." *Proceedings of NAACL-HLT*. 2016.