Neural Machine Translation is like a Pig

a.k.a NMT as an Auxiliary Task

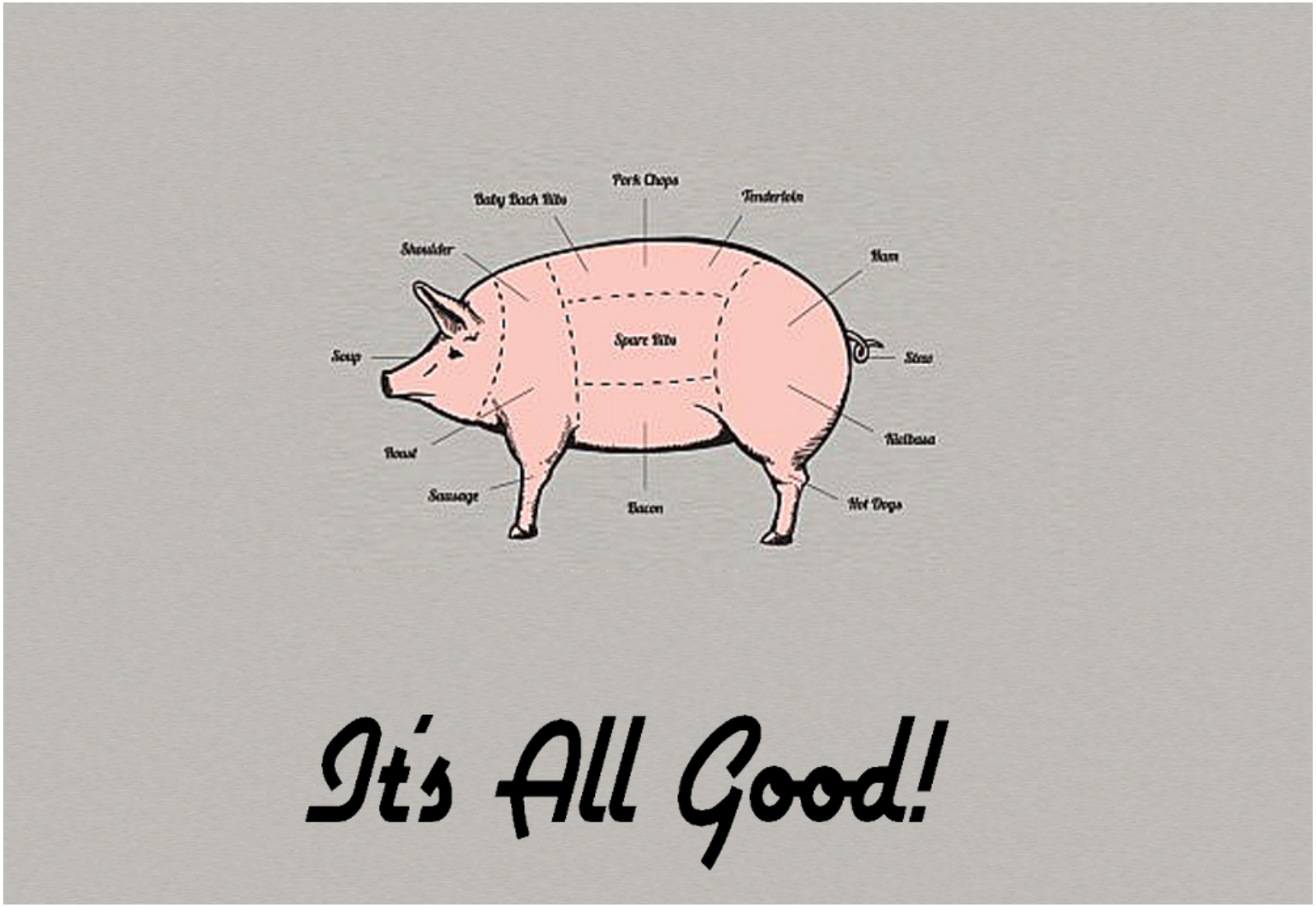
Cristina España-Bonet
UdS & DFKI

Deep Learning BCN Symposium

20th December 2018

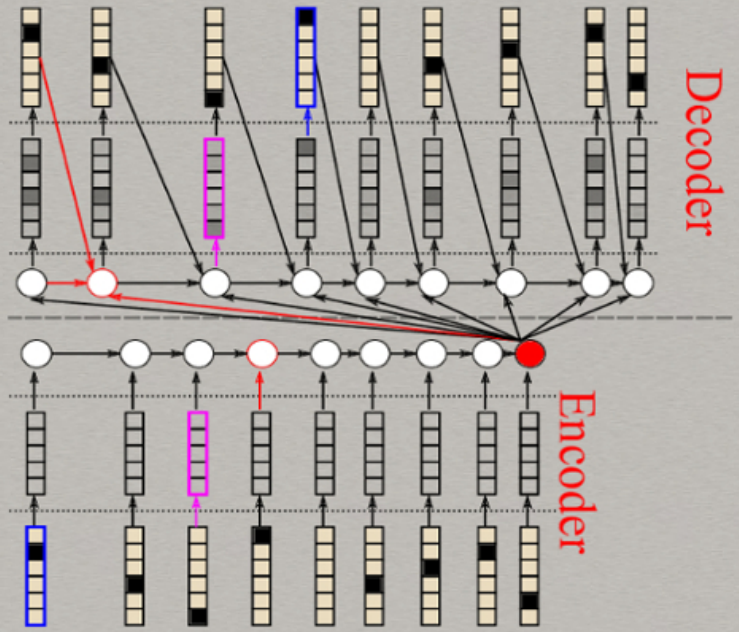
Neural Machine Translation is like a Pig

"Del porc se n'aprofita tot"



Neural Machine Translation is like a Pig

De l'NMT se n'aprofita tot!



It's All Good!

Outline

Neural Machine Translation is like a Pig

Neural Machine Translation

Seq2Seq Architecture

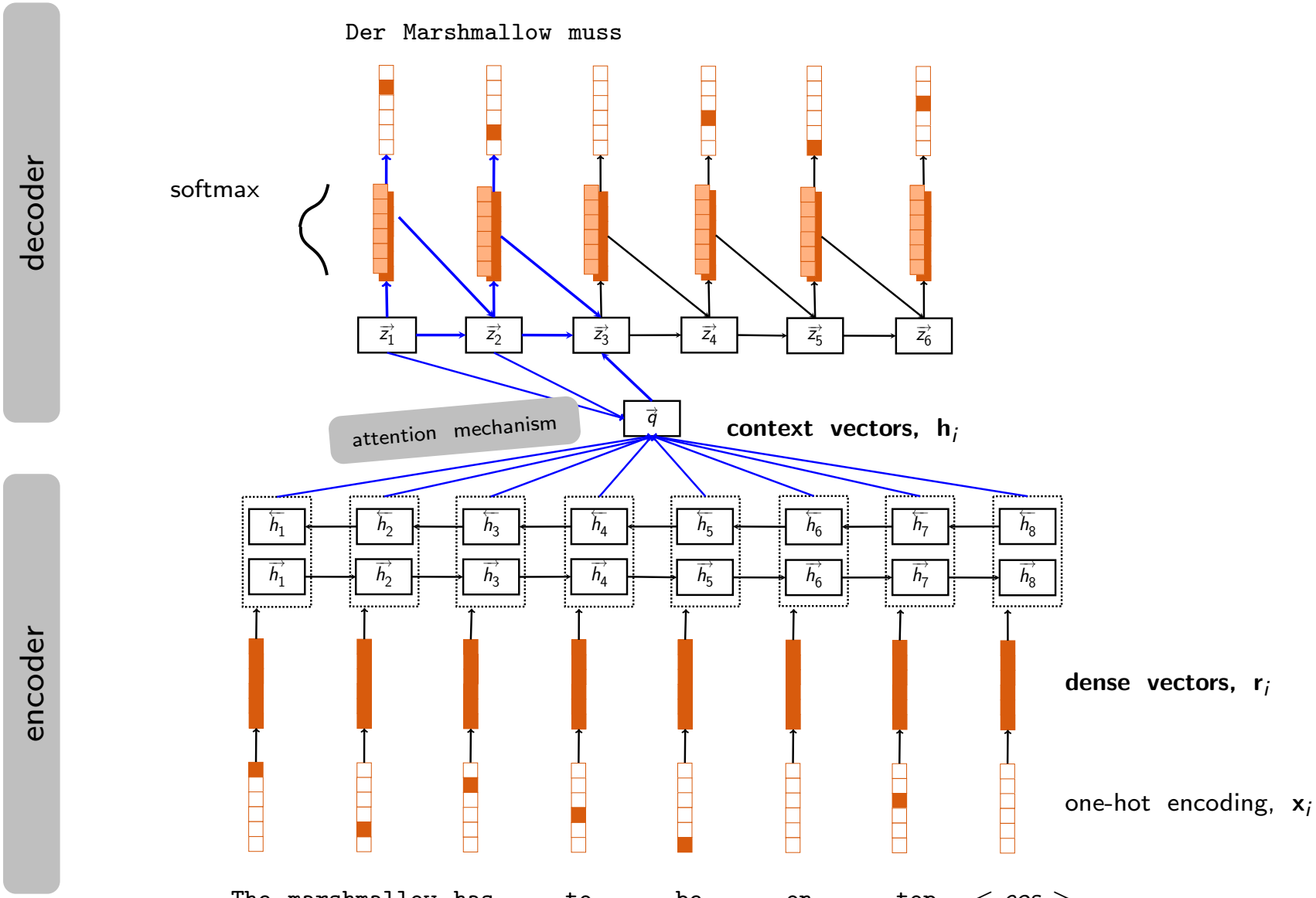
Encoder

Attention mechanism

Decoder

Neural Machine Translation

Seq2Seq Architecture



Neural Machine Translation

Seq2Seq Architecture

Encoder

Attention mechanism

Decoder

Neural Machine Translation

Seq2Seq Architecture

Encoder for source sentence $s = (x_1, \ldots, x_n)$,

$$\mathbf{r}_i = \mathbf{W_x} \cdot \mathbf{x}_i, \quad \mathbf{h}_i = \left[\overleftarrow{\mathbf{h}}_i, \overrightarrow{\mathbf{h}}_i\right] = \left[f(\overleftarrow{\mathbf{h}}_{i-1}, \mathbf{r}_i), f(\overrightarrow{\mathbf{h}}_{i+1}, \mathbf{r}_i)\right],$$

Attention mechanism

$$a(\mathbf{z}_{j-1}, \mathbf{h}_i) = \mathbf{v}_a \cdot \tanh(\mathbf{W}_a \cdot \mathbf{z}_{j-1} + \mathbf{U}_a \cdot \mathbf{h}_i)$$

$$\alpha_{ij} = \mathrm{softmax}\left(\mathrm{a}(\mathbf{z}_{\mathrm{j}-1}, \mathbf{h}_{\mathrm{i}})\right), \quad \mathbf{q}_j = \sum_i \alpha_{ij} \mathbf{h}_i$$

Decoder for target sentence $t = (y_1, \ldots, y_n)$,

$$\mathbf{z}_j = g(\mathbf{z}_{j-1}, \mathbf{t}_{j-1}, \mathbf{q}_j), \quad \mathbf{t}_{j-1} = \mathbf{W_y} \cdot \mathbf{y}_{j-1},$$

$$p(y_j | \mathbf{y}_{<j}, \mathbf{x}) = p(y_j | \mathbf{z}_j, \mathbf{t}_{j-1}, \mathbf{q}_j) = \mathrm{softmax}\left(\mathbf{p}_j \mathbf{W}_o\right),$$

$$\mathbf{p}_j = \tanh\left(\mathbf{z}_j \mathbf{W}_{p1} + \mathbf{W_y}[y_{j-1}] \mathbf{W}_{p2} + \mathbf{q}_j \mathbf{W}_{p3}\right)$$

NMT representations, why?

(Multilingual) Sentence Representations

NMT representations, why? Multilinguality!

Neural Machine Translation

(Multilingual) Sentence Representations

NMT representations, why? Multilinguality!

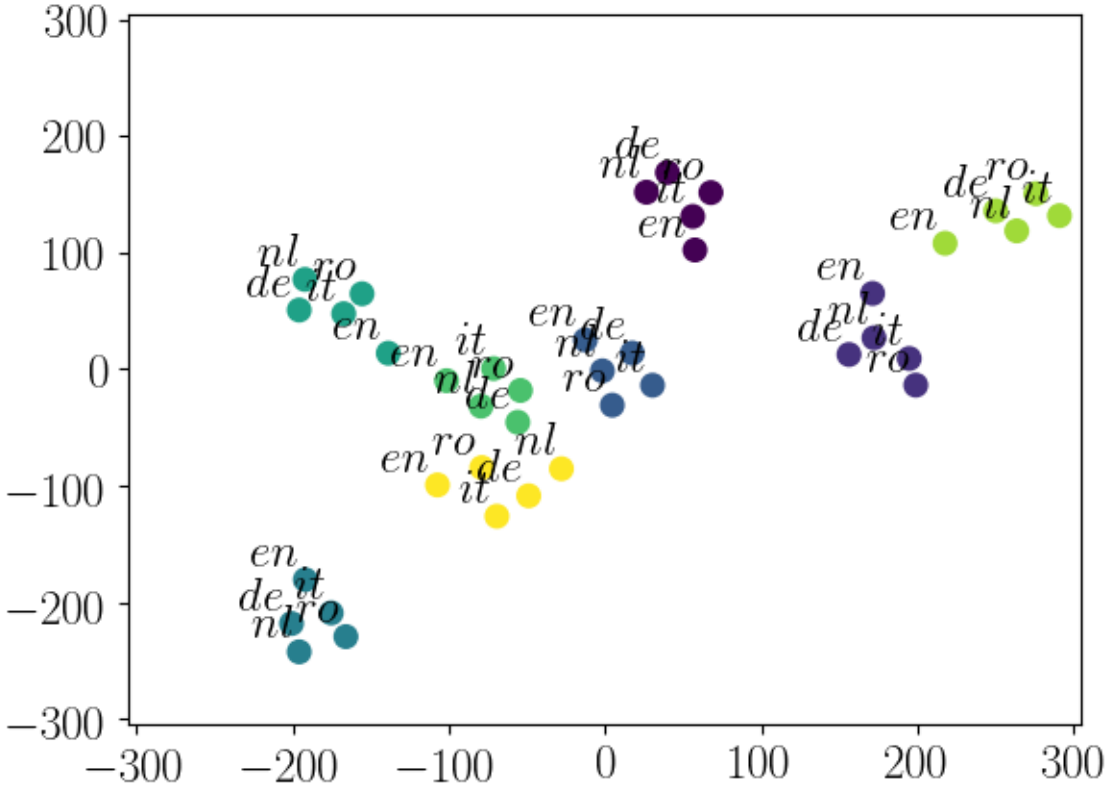
- Machine Translation is naturally a bilingual task
- Neural Machine Translation (NMT) encodes semantics in vectors
- Straightforward extension of NMT to multilingual NMT (ML-NMT)
- ML word (or context) vectors lie in the same space

Crosslingual Textual Similarity

Crosslingual Textual Similarity

Multilingual Semantic Space for Context Vectors (easy)

(España-Bonet & van Genabith, 2018)

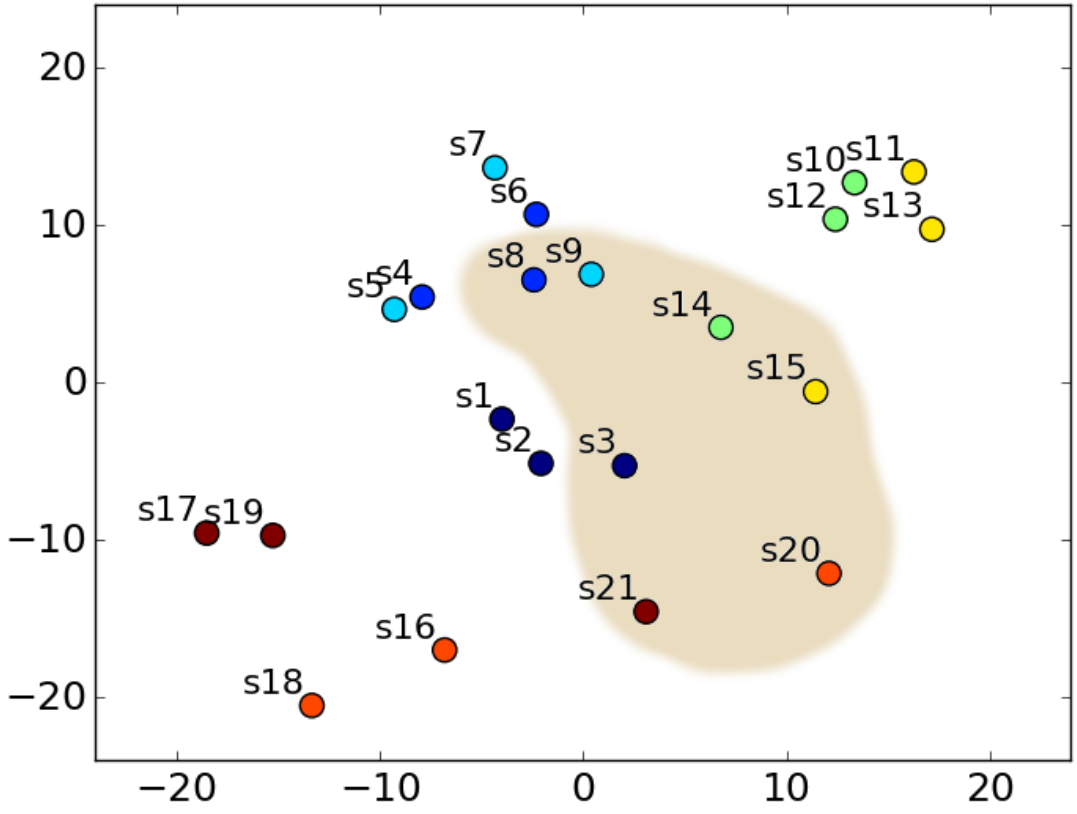


ML-NMT $\{de, en, nl, it, ro\} \rightarrow \{de, en, nl, it, ro\}$ with TED talks

Crosslingual Textual Similarity

Multilingual Semantic Space for Context Vectors (hard)

(España-Bonet et al., 2017)



ML-NMT $\{en, es, ar\} \rightarrow \{en, es, ar\}$ with heterogeneous corpora

Crosslingual Textual Similarity

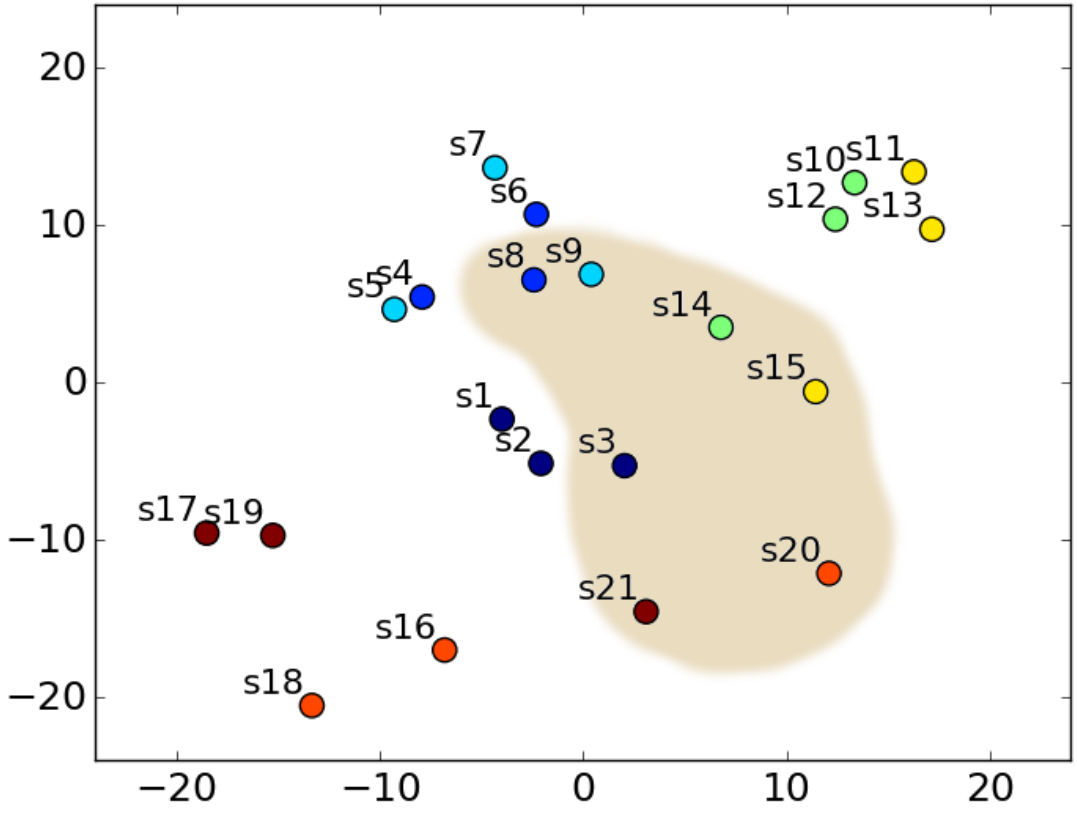
Multilingual Semantic Space for Context Vectors (hard)

$s1{:}t1$	Spain princess testifies in historic fraud probe
$s2{:}t1$	Princesa de España testifica en juicio histórico de fraude
$s3{:}t1$	أميرة أسبانيا تدلي بشهادتها في قضية احتيال تاريخي.
$s4{:}t2$	You do not need to worry.
$s5{:}t3$	You don't have to worry.
$s6{:}t2$	No necesitas preocuparte.
$s7{:}t3$	No te tienes por que preocupar.
$s8{:}t2$	لا ينبغي أن تقلق
$s9{:}t3$	لا ينبغي أن تجزع.
$s10{:}t4$	Mandela's condition has 'improved'
$s11{:}t5$	Mandela's condition has 'worsened over past 48 hours'
$s12{:}t4$	La salud de Mandela ha 'mejorado'
$s13{:}t5$	La salud de Mandela 'ha empeorado en las últimas 48 horas'
$s14{:}t4$	لقد تحسنّت حالة مانديلا الصحية.
$s15{:}t5$	ساءت الحالة الصحية لمانديلا خلال ال ٤٨ ساعة الماضية.
$s16{:}t6$	Vector space representation results in the loss of the order which the terms are in the document.
$s17{:}t7$	If a term occurs in the document, the value will be non-zero in the vector.
$s18{:}t6$	La representación en el espacio de vecores implica la pérdida del órden en el que los términos ocurren en el documento.
$s19{:}t7$	Si un término ocurre en el document, el valor en el vector será distinto de cero.
$s20{:}t6$	يؤدي تمثيلُ فضاءِ المتجهِ إلى فقد الترتيب الذي تكون عليه المصطلحات في الوثيقة.
$s21{:}t7$	إذا ما ورد مصطلح في الوثيقة، فالقيمة ستكون غيرصفرية المتجه.

Crosslingual Textual Similarity

Multilingual Semantic Space for Context Vectors (hard)

(España-Bonet et al., 2017)

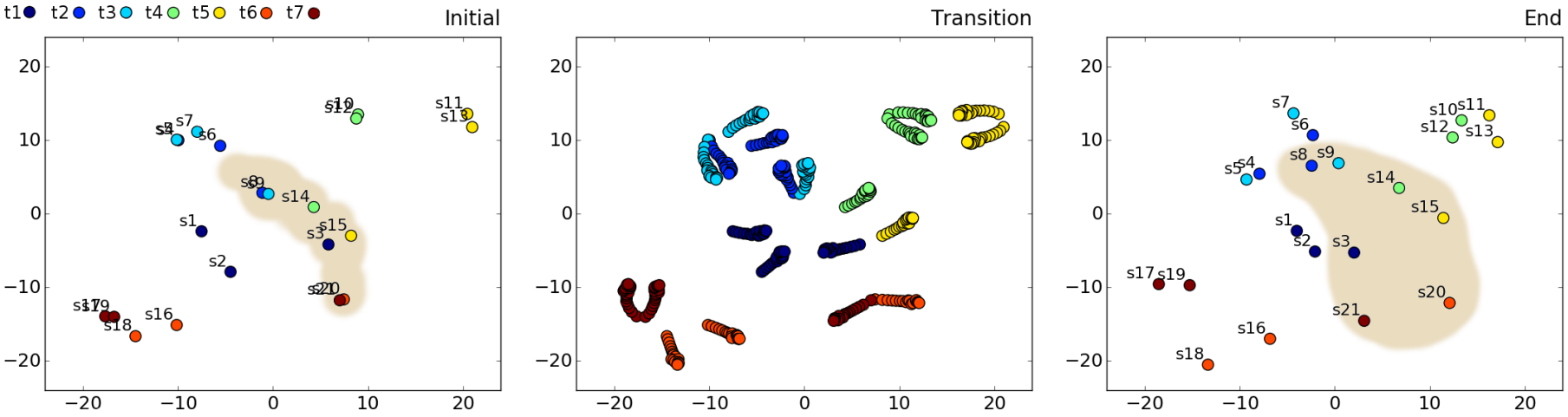


ML-NMT $\{en, es, ar\} \rightarrow \{en, es, ar\}$ with heterogeneous corpora

Crosslingual Textual Similarity

Evolution of Context Vectors through Training (hard)

(España-Bonet et al., 2017)



ML-NMT $\{en,\ es,\ ar\} \rightarrow \{en,\ es,\ ar\}$ with heterogeneous corpora

Crosslingual Textual Similarity

Evolution of Context Vectors through Training (hard)

Pearson correlation on the Semantic Textual Similarity Task
STS 2017 data

	track1 *ar–ar*	track2 *ar–en*	track3 *es–es*	track4a *es–en*	track5 *en–en*
`WE-d300`	0.49	0.28	0.55	0.40	0.56
`WE-d1024`	0.51	0.33	0.59	0.45	0.60

(España-Bonet & Barrón-Cedeño, 2017)

Crosslingual Textual Similarity

Evolution of Context Vectors through Training (hard)

Pearson correlation on the Semantic Textual Similarity Task
STS 2017 data

	track1 *ar–ar*	track2 *ar–en*	track3 *es–es*	track4a *es–en*	track5 *en–en*
`WE-d300`	0.49	0.28	0.55	0.40	0.56
`WE-d1024`	0.51	0.33	0.59	0.45	0.60
$\texttt{NMT}_{\text{ctx}}$-0.1Ep	0.32	0.25	0.55	0.32	0.54
$\texttt{NMT}_{\text{ctx}}$-0.5Ep	0.52	0.36	0.71	0.40	0.68
$\texttt{NMT}_{\text{ctx}}$-1.0Ep	0.57	0.42	0.74	0.44	0.72
$\texttt{NMT}_{\text{ctx}}$-2.0Ep	0.59	0.44	0.78	0.49	0.76

(España-Bonet & Barrón-Cedeño, 2017)

Crosslingual Textual Similarity

Cool Related Research: Devlin et al., 2018

Pearson correlation on the Semantic Textual Similarity Task

	track1 *ar–ar*	track2 *ar–en*	track3 *es–es*	track4a *es–en*	track5 *en–en*
`WE-d300`	0.49	0.28	0.55	0.40	0.56
`WE-d1024`	0.51	0.33	0.59	0.45	0.60
$\mathtt{NMT}_{\mathrm{ctx}}$-2.0Ep	0.59	0.44	0.78	0.49	0.76
BERT	?	?	?	?	*0.59*

Bidirectional **E**ncoder **R**epresentations from **T**ransformers
Devlin *et al.*, 2018
Google AI Language

Crosslingual Textual Similarity

Cool Related Research: Devlin et al., 2018

Pearson correlation on the Semantic Textual Similarity Task

	track1 *ar–ar*	track2 *ar–en*	track3 *es–es*	track4a *es–en*	track5 *en–en*
`WE-d300`	0.49	0.28	0.55	0.40	0.56
`WE-d1024`	0.51	0.33	0.59	0.45	0.60
$\mathtt{NMT}_{\mathrm{ctx}}$-2.0Ep	0.59	0.44	0.78	0.49	0.76
BERT	?	?	?	?	*0.59*
BERT+adapt	?	?	?	?	0.85
BERT$_{LARGE}$+adapt	?	?	?	?	0.86

Crosslingual Textual Similarity

Cool Related Research: Bert Embeddings

method	PPMCC (STS-B dev)
bert, no FT, cosine similarity between sentence embedding (`[CLS]`)	0.29
bert, FT, simple regression	0.89
bert, FT, cosine similarity between sentence embedding (`[CLS]`)	0.66
bert, no FT, cosine similarity between mean-pooled sequence embeddings (`mean_pool([CLS], tok1, ..., [SEP])`)	0.59
average word vector (spaCy, `en_core_web_lg`)	0.54

Crosslingual Textual Similarity

What else?

Parallel sentence extraction, as a pre-process or on-line

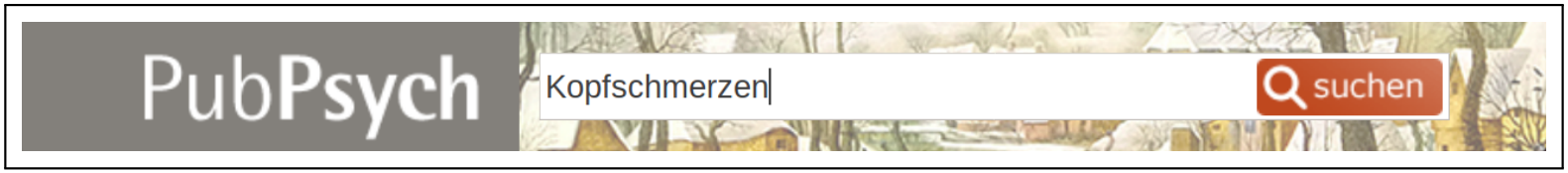
	de–en			*fr–en*			joint		
	P	R	F_1	P	R	F_1	P	R	F_1
Thrs.	95.5	97.1	96.3	95.4	100.0	97.7	98.3	98.1	98.2
SVM	96.2	96.2	96.2	95.6	99.1	97.3	97.1	98.0	97.6
GB	97.0	95.7	96.4	95.6	99.6	97.6	97.0	97.3	97.2
Ens.	98.2	95.7	97.0	95.6	99.1	97.3	96.9	97.8	97.3

(on data of the BUCC 2017 shared task)

Multilingual Query Expansion

Multilingual Query Expansion

Motivation

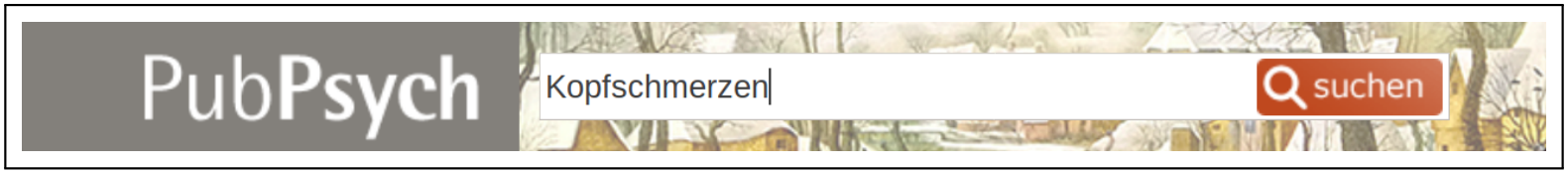


Parsed query: +(Kopfschmerzen)

String components of query: Kopfschmerzen

Multilingual Query Expansion

Motivation



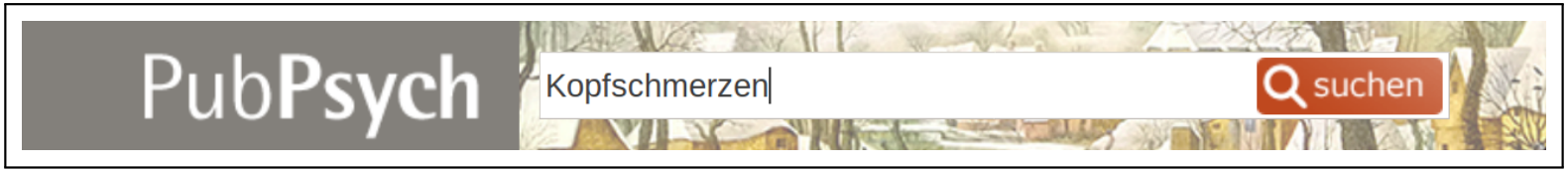
Parsed query: +(Kopfschmerzen)

String components of query: **Kopfschmerzen**

Translations:
cp::Kopfschmerzen

Multilingual Query Expansion

Motivation



Parsed query: +(Kopfschmerzen)

String components of query: **Kopf@@, schmerzen**

Translations:

en::head es::cabeza fr::tête en::ache es::dolor fr::douleur

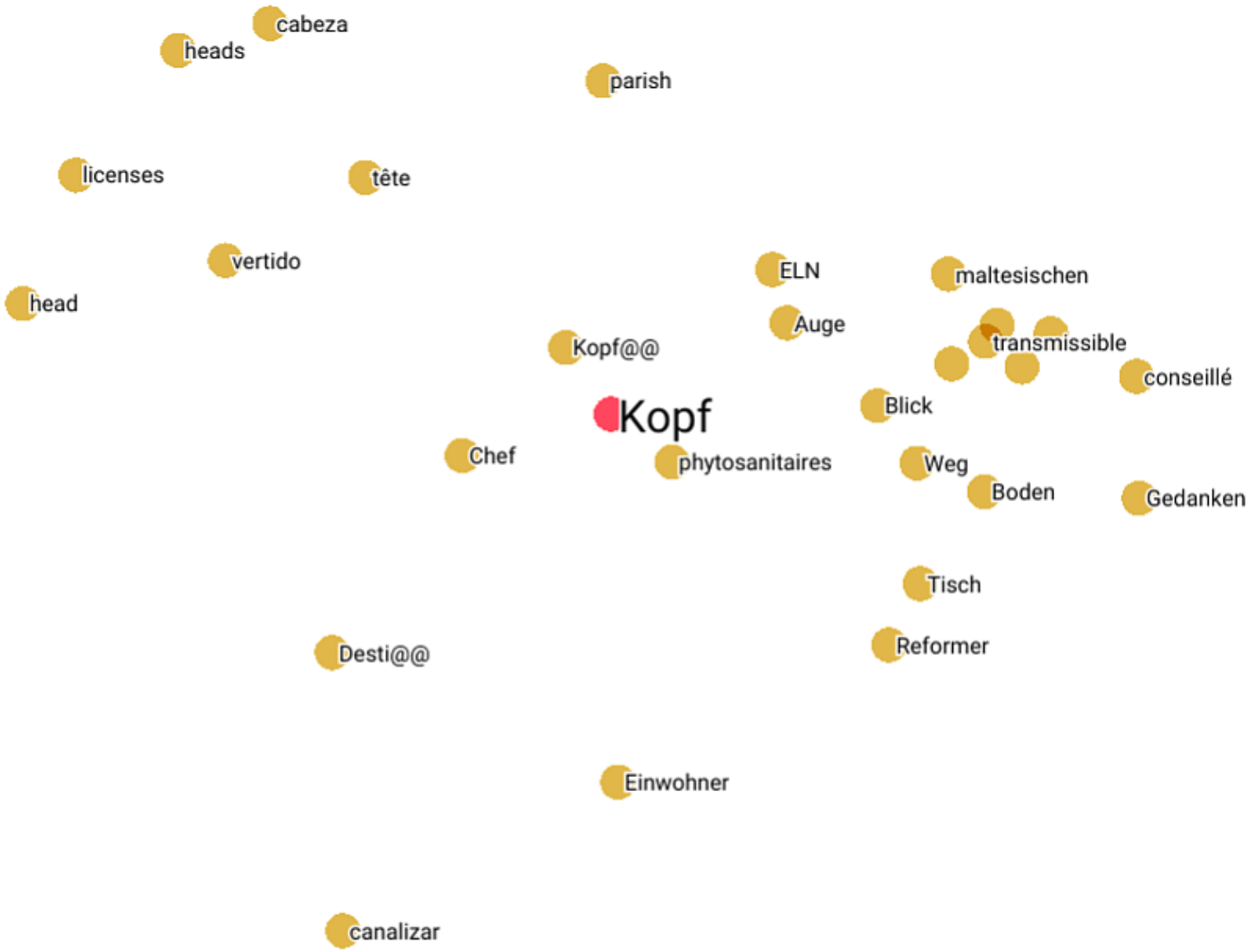
Multilingual Query Expansion

ML-NMT

- ML-NMT $\{en, es, fr, de\} \Rightarrow \{en, es, fr, de\}$
- Joint 120K vocabulary (+BPE)
- General Domain 96,000,000 parallel sentences (balanced)
- Seq2seq: 3 months of training
- Transformer: 1 month of training

Multilingual Query Expansion

ML-NMT Word Embeddings, 2D Representation



Multilingual Query Expansion

ML-NMT Word Embeddings, 2D Representation

Multilingual Query Expansion

Comments

- **Word** embeddings do not deal with multiword expressions or compounds
- Embeddings describe **semantics** not language, a synonym or a translation or a synonym of a translation have the same right to be close to a given word
  - **Query expansion** or query translation?
- Even within a language, translations do not need to be the **closest word** $\Rightarrow$ **Need for reranking**

Multilingual Query Expansion

Seq2seq ML Word Embeddings

Kopf vs. think

Nearest points in the original space:

transmissible	0.606
cabeza ★	0.618
head ★	0.619
heads	0.621
Kopf@@	0.631
conseillé	0.635
Desti@@	0.638
tête ★	0.657
maltesischen	0.664
capita	0.666

Multilingual Query Expansion

Seq2seq ML Word Embeddings

Kopf vs. think

Nearest points in the original space:

transmissible	0.606
cabeza ★	0.618
head ★	0.619
heads	0.621
Kopf@@	0.631
conseillé	0.635
Desti@@	0.638
tête ★	0.657
maltesischen	0.664
capita	0.666

Nearest points in the original space:

believe	0.193
consider	0.416
thinks	0.432
thought	0.432
feel	0.443
believes	0.492
thinking	0.522
know	0.531
pense ★	0.539
believed	0.549

Multilingual Query Expansion

Seq2seq ML Word Embeddings

Kopf vs. think

Nearest points in the original space:

transmissible		0.606
cabeza	★	0.618
head	★	0.619
heads		0.621
Kopf@@		0.631
conseillé		0.635
Desti@@		0.638
tête	★	0.657
maltesischen		0.664
capita		0.666

Nearest points in the original space:

believe		0.193
consider		0.416
thinks		0.432
thought		0.432
feel		0.443
believes		0.492
thinking		0.522
know		0.531
pense	★	0.539
believed		0.549

pensar	★	24 th	0.606
denken	★	56 th	0.670

Multilingual Query Expansion

Transformer ML Word Embeddings

Kopf vs. think

Nearest points in the original space:

cabeza	★	0.841
tête	★	0.860
capita		0.921
Kopf-@@		0.929
habitant		0.944
head	★	0.947
kopf		0.985
Pro-Kopf-@@		0.985
cápita		0.987
habitante		0.994
crâ@@		1.019

Nearest points in the original space:

thinks		0.796
thought		0.832
believe		0.833
thinking		0.840
pense	★	0.933
denken	★	0.975
Meinung		0.983
penser		0.992
pensent		1.003
denke		1.007
believes		1.008
pensar	★	1.017
creo		1.026
pensons		1.028

Conclusions

Summary

- ML-NMT construct multilingual/interlingual semantic spaces for words and sentences, with different characteristics for seq2seq and transformers
- NMT word and sentence context vectors are useful representations for several ML-NLP tasks

Conclusions

Summary

- ML-NMT construct multilingual/interlingual semantic spaces for words and sentences, with different characteristics for seq2seq and transformers
- NMT word and sentence context vectors are useful representations for several ML-NLP tasks
- In our research we succesfully use them for:
  * Semantic similarity assessments
  * Parallel sentence extraction
  * On-line MT training with comparable corpora
  * ML query expansion

Thanks! But...

wait!

Neural Machine Translation is like a Pig

a.k.a NMT as an Auxiliary Task

Cristina España-Bonet
UdS & DFKI

Deep Learning BCN Symposium

20th December 2018

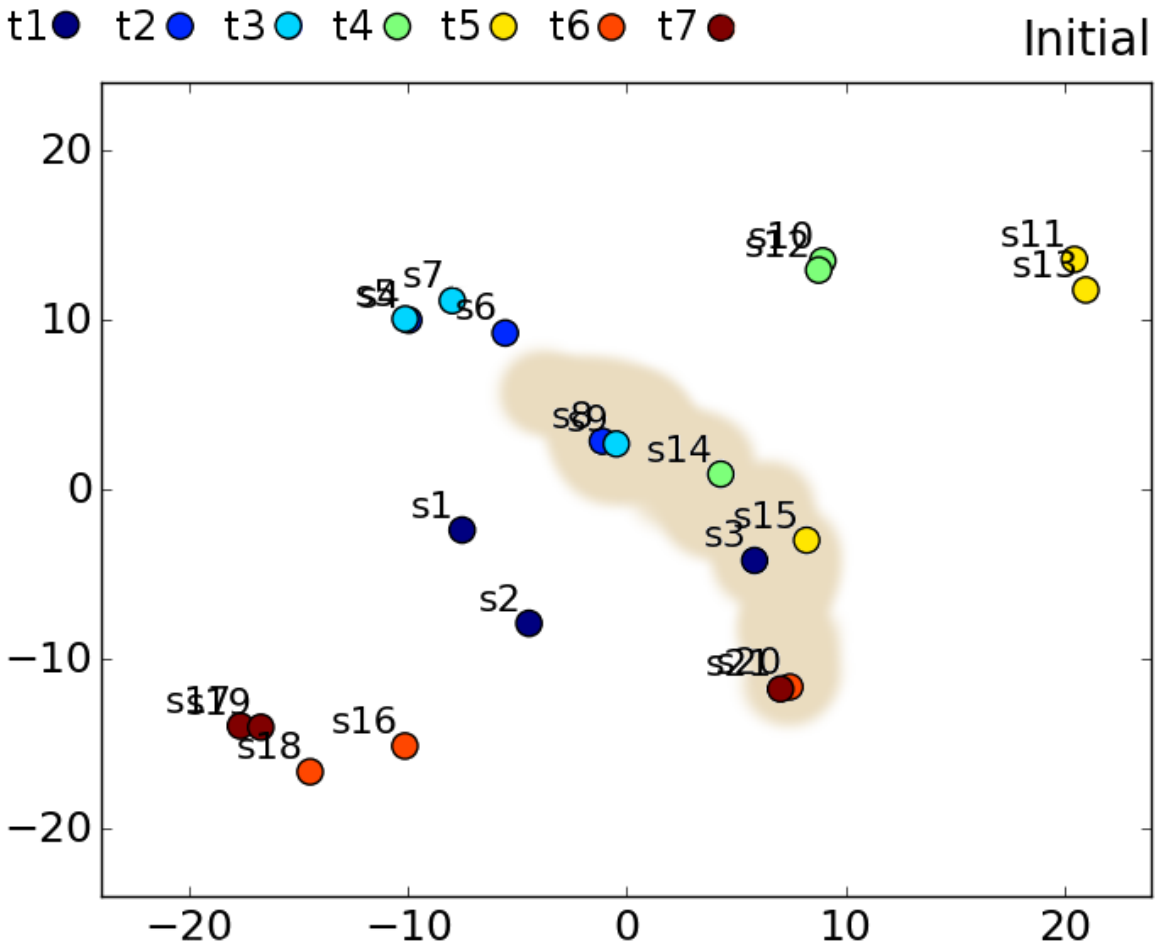
Extra Slides: ML Sentence Embeddings

Evolution of Context Vectors through Training II

$s1{:}t1$	Spain princess testifies in historic fraud probe
$s2{:}t1$	Princesa de España testifica en juicio histórico de fraude
$s3{:}t1$	أميرة أسبانيا تدلي بشهادتها في قضية احتيال تاريخي.
$s4{:}t2$	You do not need to worry.
$s5{:}t3$	You don't have to worry.
$s6{:}t2$	No necesitas preocuparte.
$s7{:}t3$	No te tienes por que preocupar.
$s8{:}t2$	لا ينبغي أن تقلق
$s9{:}t3$	لا ينبغي أن تجزع.
$s10{:}t4$	Mandela's condition has 'improved'
$s11{:}t5$	Mandela's condition has 'worsened over past 48 hours'
$s12{:}t4$	La salud de Mandela ha 'mejorado'
$s13{:}t5$	La salud de Mandela 'ha empeorado en las últimas 48 horas'
$s14{:}t4$	لقد تحسنّت حالة مانديلا الصحية.
$s15{:}t5$	ساءت الحالة الصحية لمانديلا خلال ال ٤٨ ساعة الماضية.
$s16{:}t6$	Vector space representation results in the loss of the order which the terms are in the document.
$s17{:}t7$	If a term occurs in the document, the value will be non-zero in the vector.
$s18{:}t6$	La representación en el espacio de vecores implica la pérdida del órden en el que los términos ocurren en el documento.
$s19{:}t7$	Si un término ocurre en el document, el valor en el vector será distinto de cero.
$s20{:}t6$	يؤدي تمثيلُ فضاءِ المتجهِ إلى فقد الترتيب الذي تكون عليه المصطلحات في الوثيقة.
$s21{:}t7$	إذا ما ورد مصطلح في الوثيقة، فالقيمة ستكون غيرصفرية المتجه.

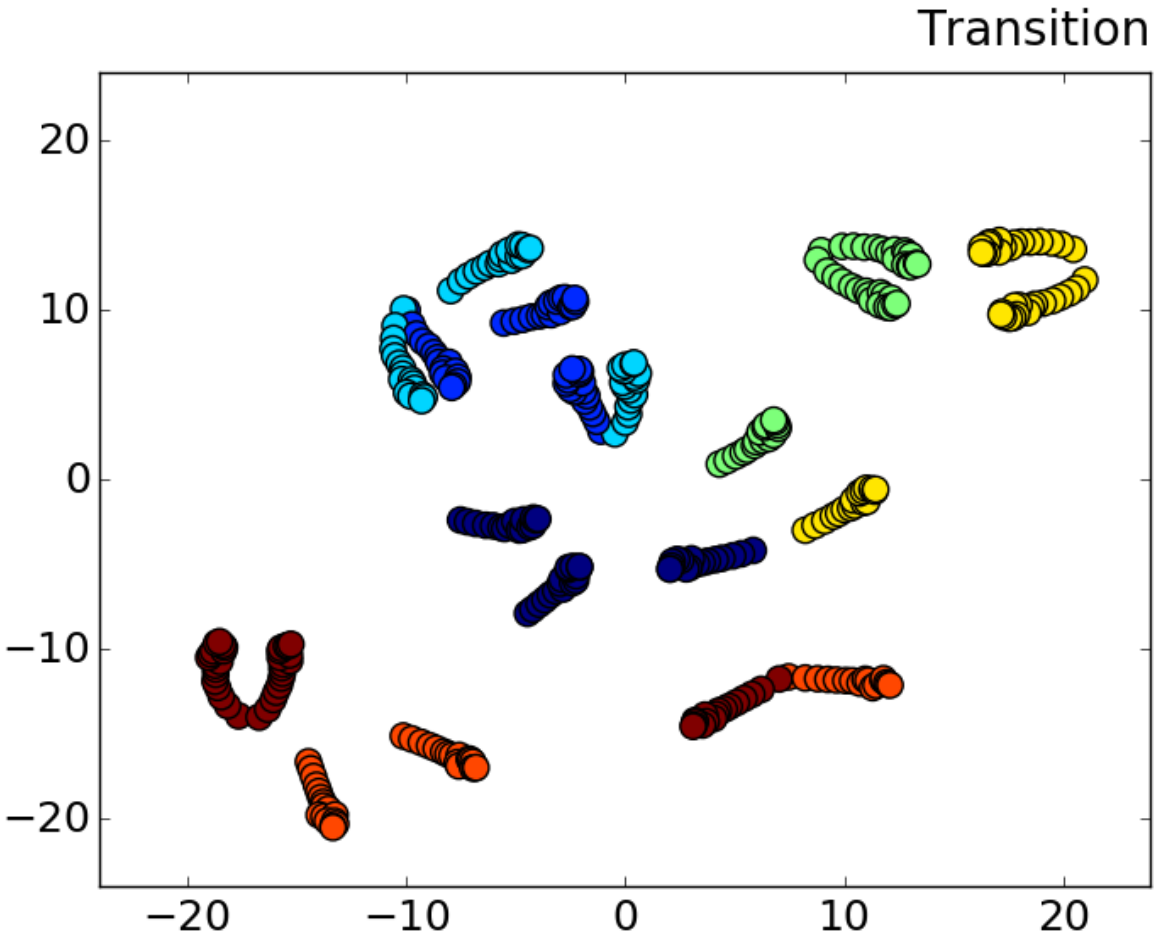
Extra Slides: ML Sentence Embeddings

Evolution of Context Vectors through Training III



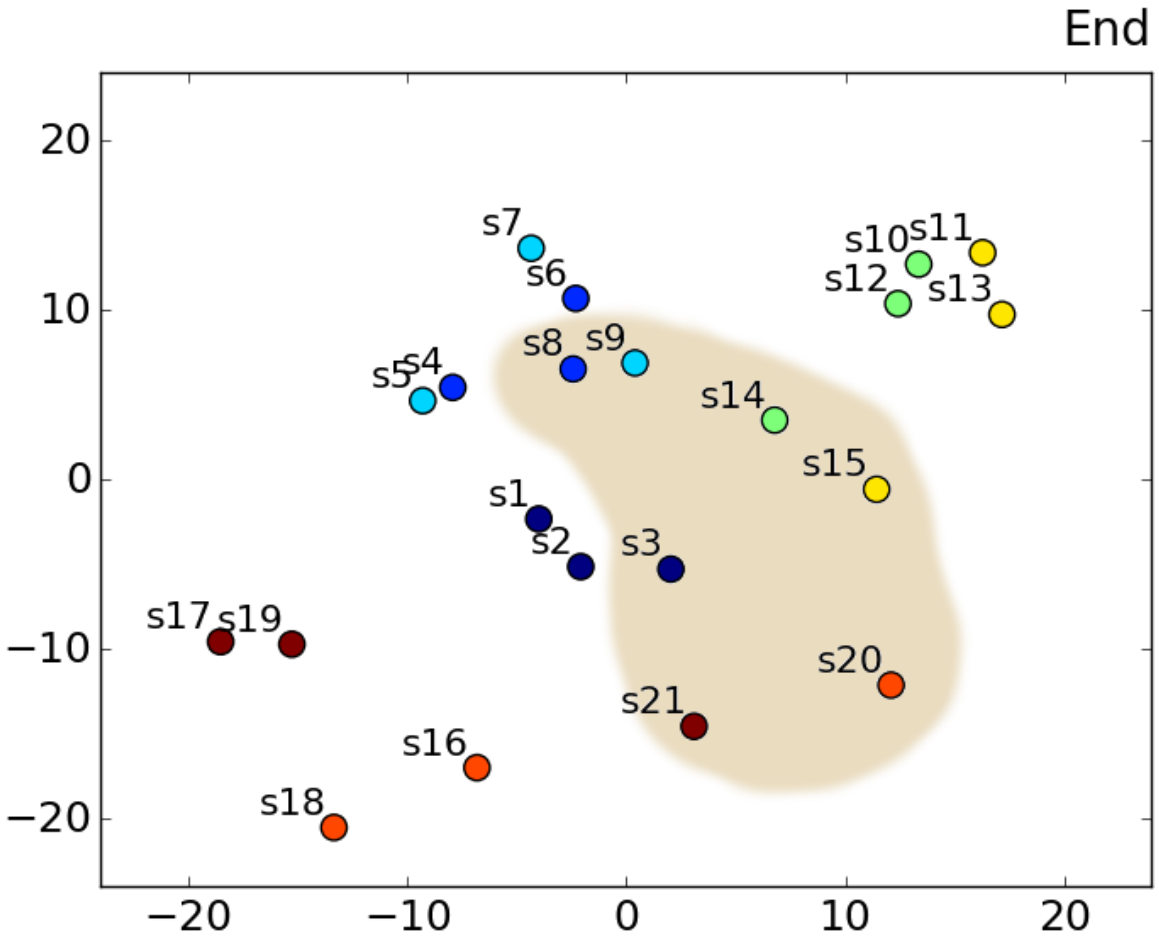
Extra Slides: ML Sentence Embeddings

Evolution of Context Vectors through Training III



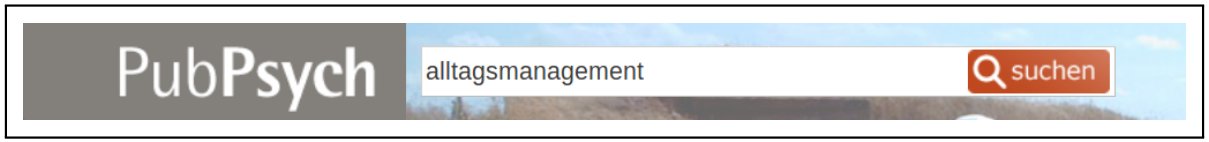
Extra Slides: ML Sentence Embeddings

Evolution of Context Vectors through Training III



Extra Slides: ML Word Embeddings

Query Translation Limitations

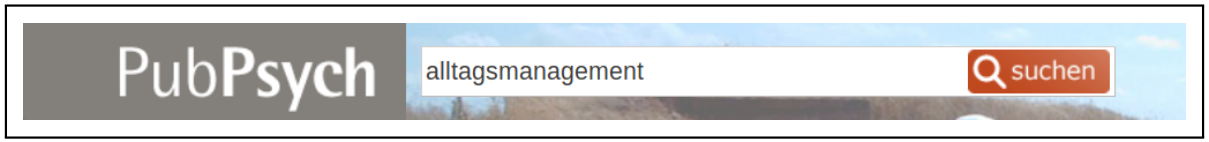


Parsed query: +(alltagsmanagement)

String components of query: alltagsmanagement

Extra Slides: ML Word Embeddings

Query Translation Limitations



Parsed query: +(alltagsmanagement)

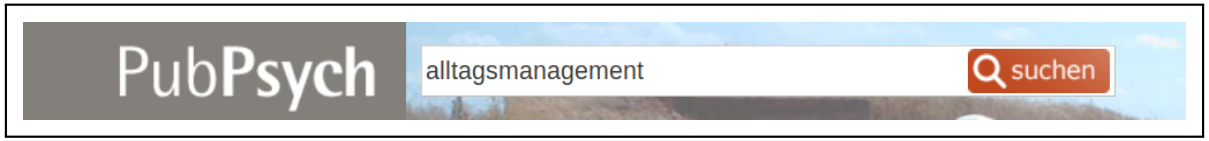
String components of query: **alltagsmanagement**

Translations:

cp::alltagsmanagement

Extra Slides: ML Word Embeddings

Query Translation Limitations

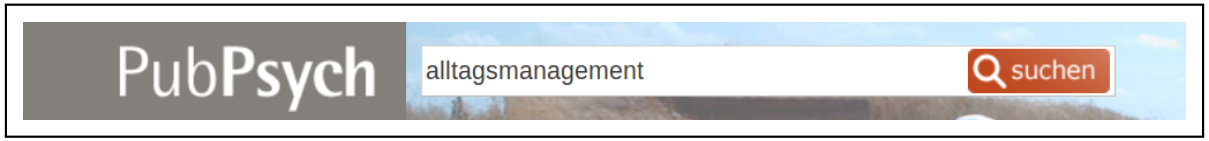


Parsed query: +(alltagsmanagement)

String components of query: alltags@@, management

Extra Slides: ML Word Embeddings

Query Translation Limitations



Parsed query: +(alltagsmanagement)

String components of query: **alltags@@, management**

Translations:

en::everyday es::cotidiano fr::ordinaires

en::management es::gestión fr::gestion

Extra Slides: ML Word Embeddings

ML-NMT Word Embeddings, 2D Representation

Extra Slides: ML Word Embeddings

ML-NMT Word Embeddings, 2D Representation

Extra Slides: ML Word Embeddings

ML-NMT Word Embeddings, 2D Representation

Extra Slides: ML Word Embeddings

Reranker

Input

Pairs source word w1 and n-best list of translations w2

Dataset

BPEd Quad-lexicon

Features

22 features: basic, lexical, semantic and LM-like

Algorithms

Mainly SVM, XGB and MLP

Extra Slides: ML Word Embeddings

Features I

Basic Features

- 2 **words** (w1 –source–, w2 –target–)
- 2 **languages** (L1, L2)
- 2 **subunits**: source is a BPE subunit instead of word (srcSubUnit), or src and tgt have a BPE mark (bothBPEmark)

Extra Slides: ML Word Embeddings

Features II

Lexical Features

- **3** **length** in characters without BPE mark @@ (l1,l2,l1/l2)
- **1** **Levenshtein distance** between tokens w1 and w2 (lev)
- **3** **character n-gram** similarity (cosSimN2,cosSimN3,cosSimN4)
- **1** Levenshtein distance between Metaphone 2 **phonetic keys** (levM2)

Extra Slides: ML Word Embeddings

Features III

Semantic Features

1. Similarity between **words**
   sim(w1, w2)
1. **Rank** of the translation, rankW2
4. **Distances** in similarities **to certain ranks**
   $sim(w1, top1) - sim(w1, w2)$
   (simRankt1,simRankWnext,simRankt10,simRankt100)

Extra Slides: ML Word Embeddings

Features IV

Language Model-like Features

1. Similarity between **previous words**
   $sim(\mathrm{w1}_{-1}, \mathrm{w2}_{-1})$

1. Similarity between **bigrams**
   $sim((\mathrm{w1}_{-1}+\mathrm{w1})/2, (\mathrm{w2}_{-1} + w2)/2)$

Extra Slides: ML Word Embeddings

Dataset: BPEd Quad-lexicon

Need for positive and negative examples:

- Positive: the translation w2
- Negative: a word w3 in the n-best list **close** to w2

Extra Slides: ML Word Embeddings

Dataset: BPEd Quad-lexicon

Need for positive and negative examples:

- Positive: the translation w2
- Negative: a word w3 in the n-best list **close** to w2

Current approach takes as w3 the +1/-1 word in the cosine similarity ranking

- For top1 translations always +1;
- higher probability to -1's to get a balanced corpus (5/6!)

Extra Slides: ML Word Embeddings

Dataset: BPEd Quad-lexicon

Need for positive and negative examples:

- Positive: the translation w2
- Negative: a word w3 in the n-best list **close** to w2

Current approach takes as w3 the +1/-1 word in the cosine similarity ranking

- For top1 translations always +1;
- higher probability to -1's to get a balanced corpus (5/6!)

Probably good for margin-based algorithms, but for others, should it be **completely random**?