Multilingual Natural Language Processing

Cristina España-Bonet

RICOH Institute of ICT, Tokyo, Japan

11th December 2017

Outline

Multilingual NLP

Semantic Representations

(Monolingual) Sentence Representations

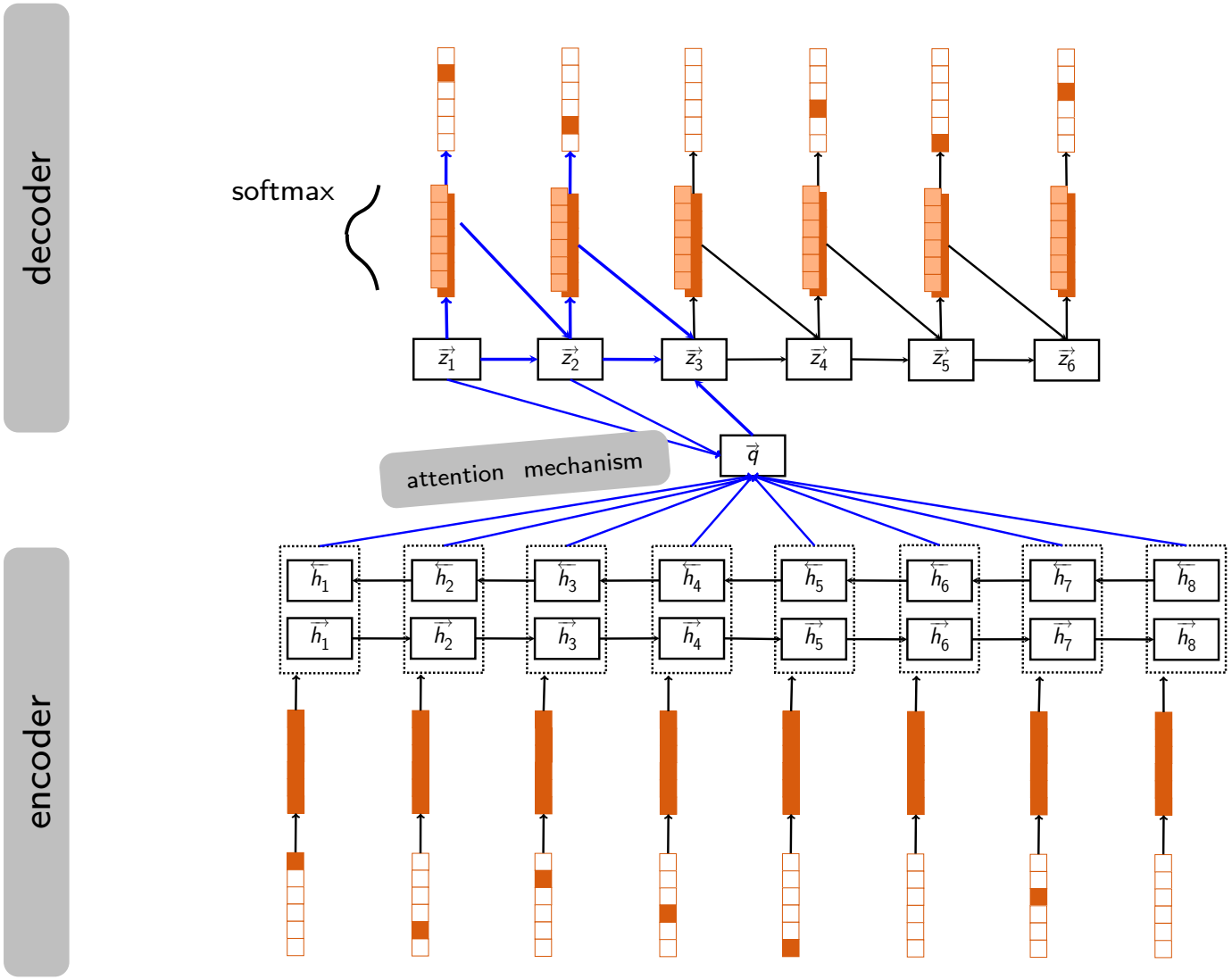
- **Composition** of word embeddings using operations $(+, \times)$ on vectors and matrices
- Internal representations in **seq2seq** architectures or auto-encoders (NMT context vectors, skip-thought vectors...)
- **Latent paragraph vectors** in word2vec-like NNs

ML-NMT Context Vectors, why?

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

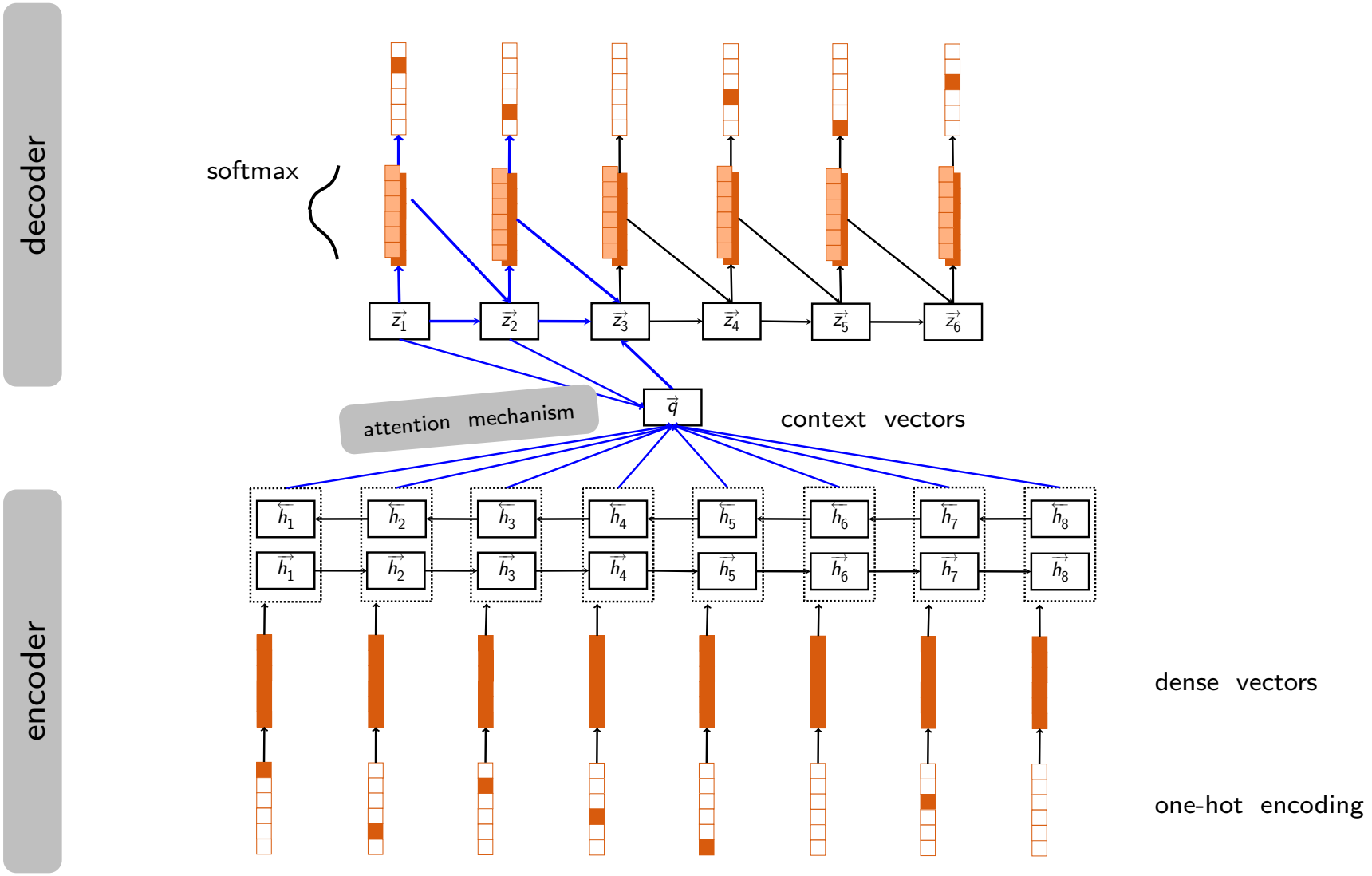
Semantic Representations

in Neural Machine Translation Architectures



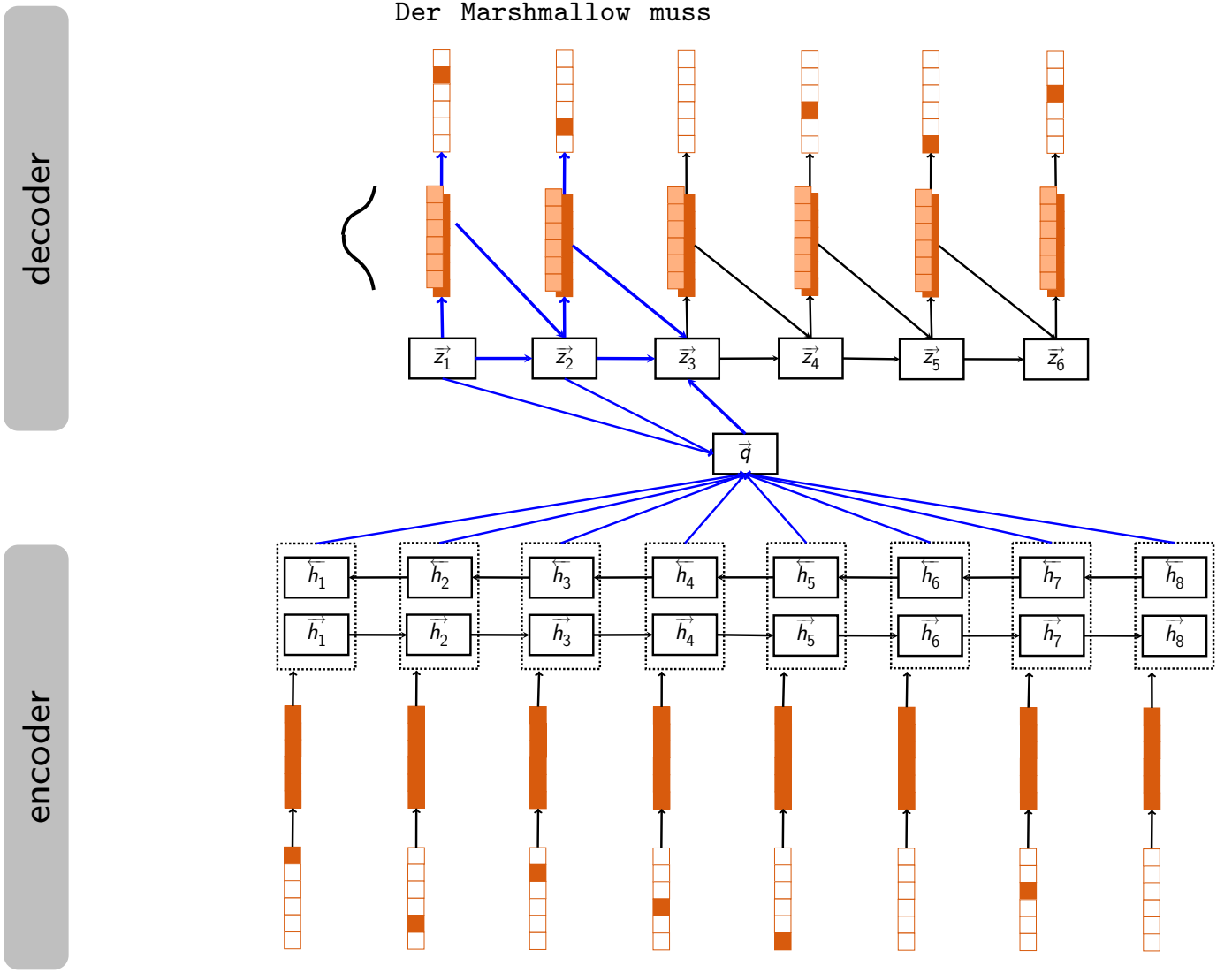
Semantic Representations

in Neural Machine Translation Architectures



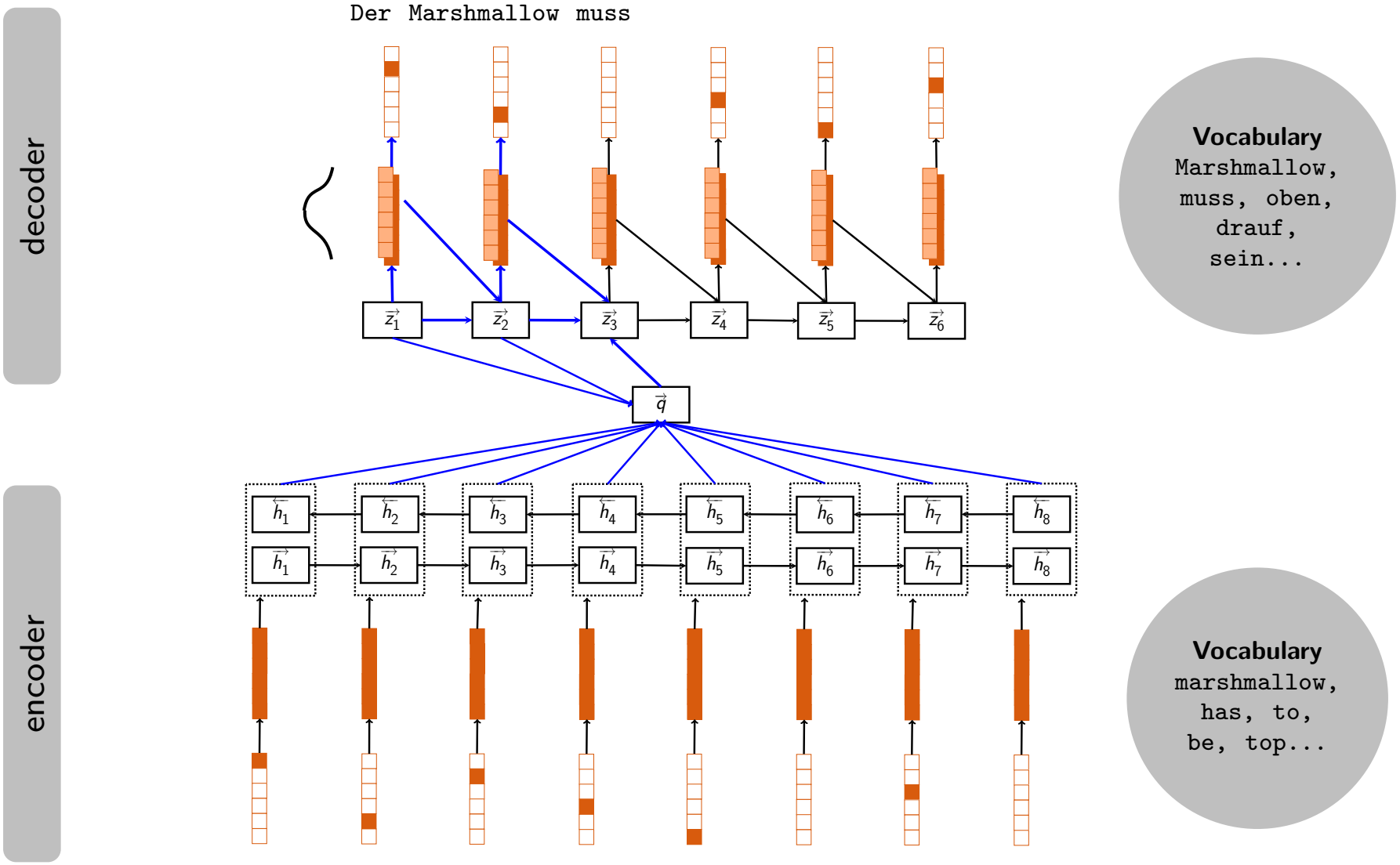
Semantic Representations

in Neural Machine Translation Architectures



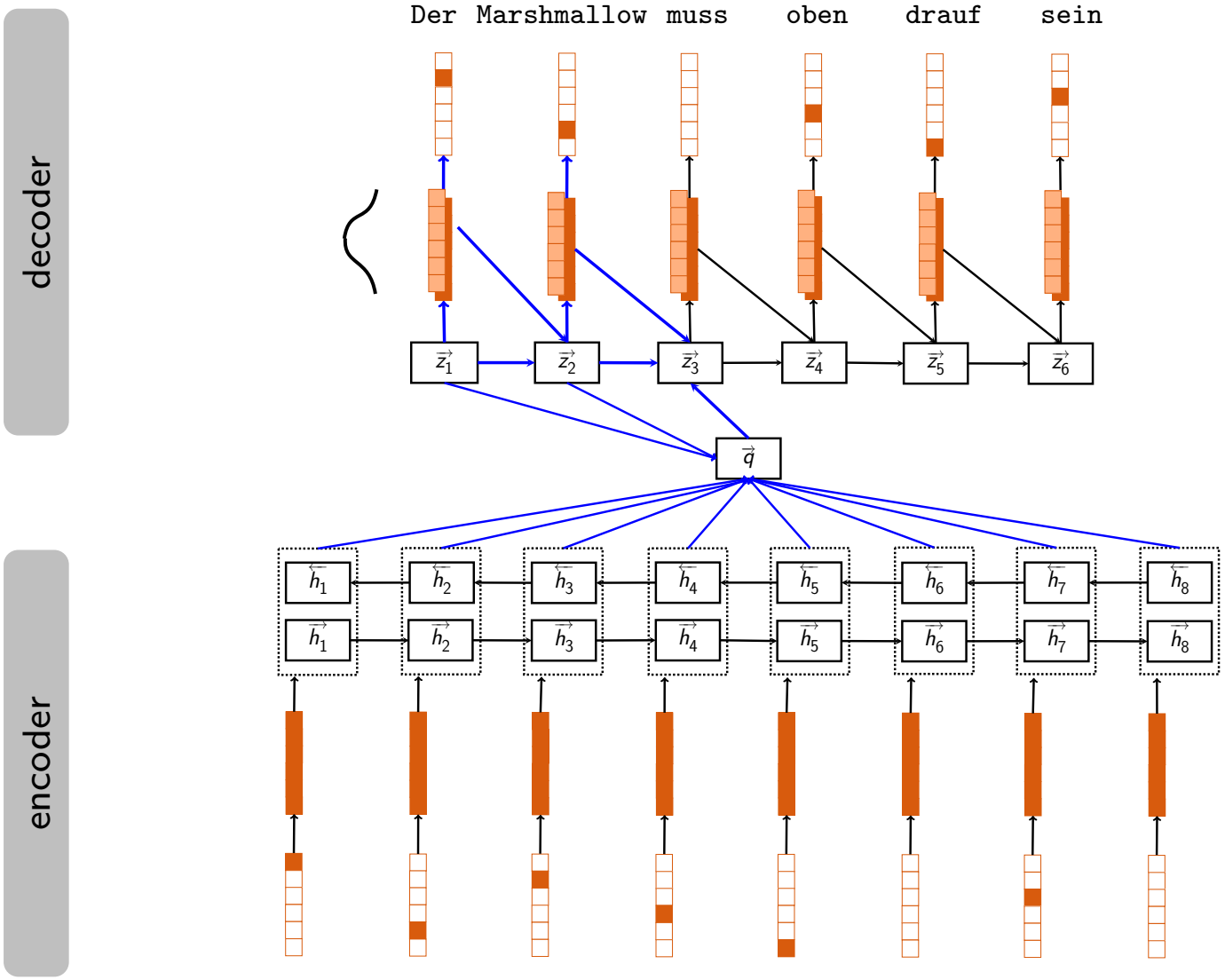
Semantic Representations

in Neural Machine Translation Architectures



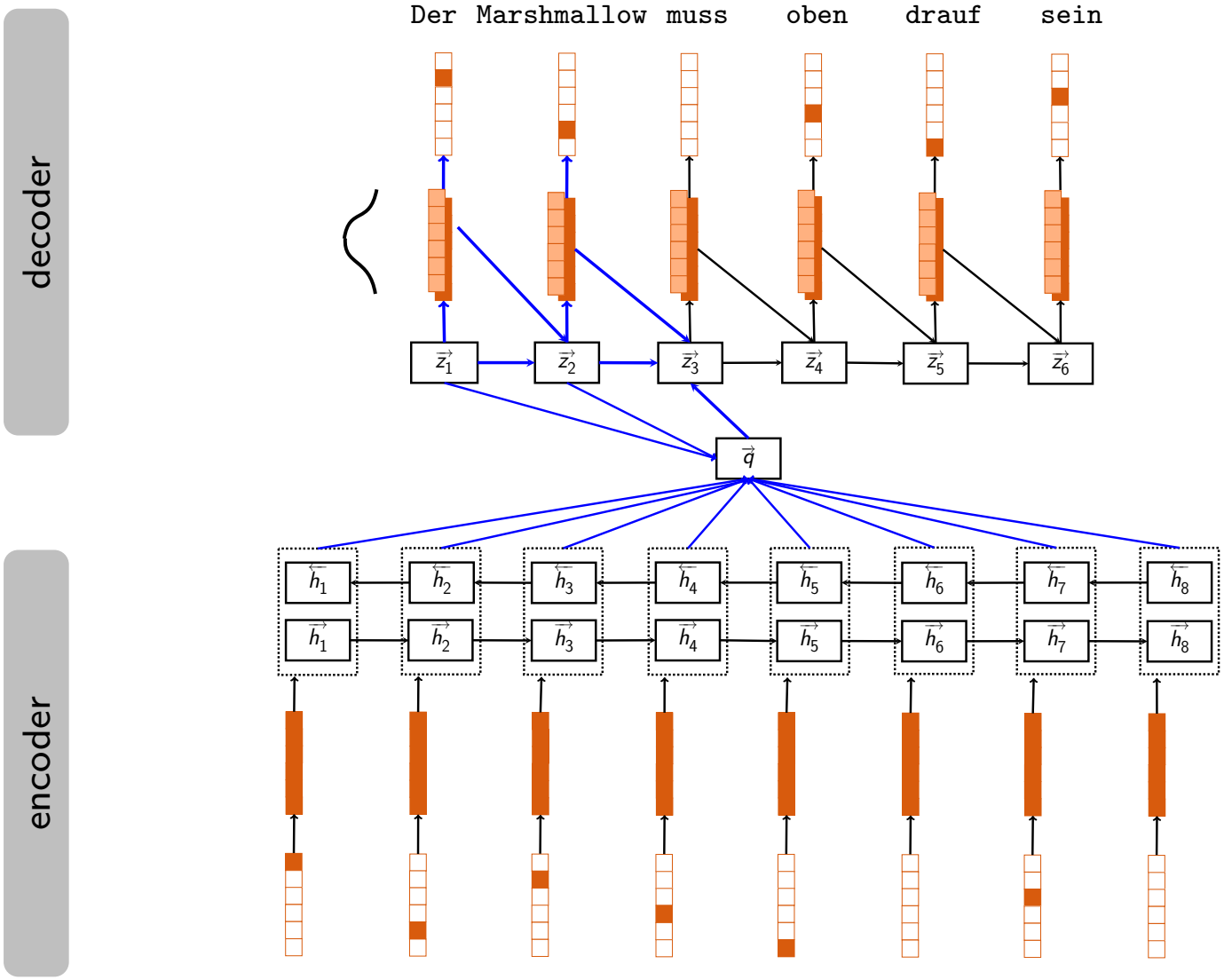
Semantic Representations

in Neural Machine Translation Architectures



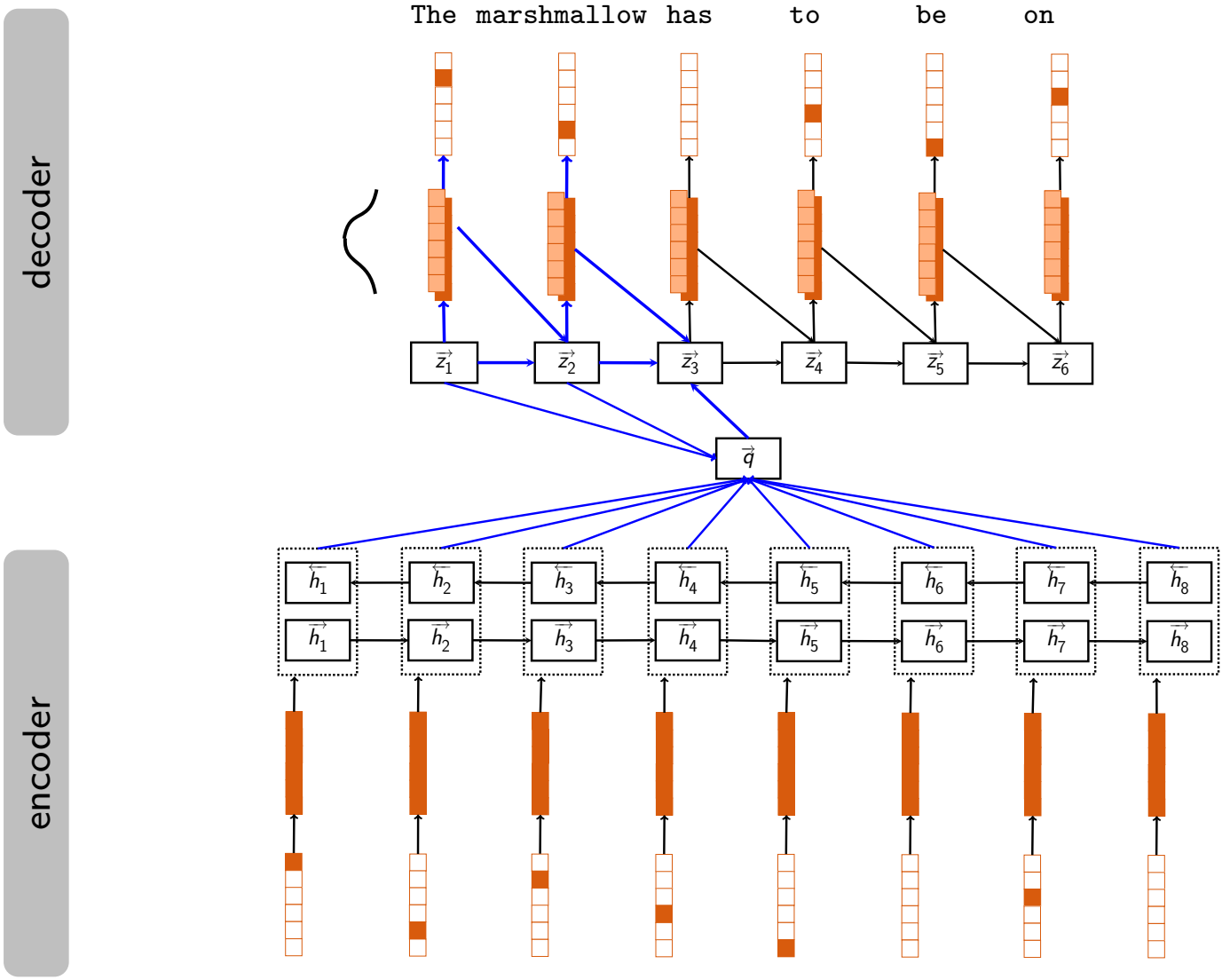
Semantic Representations

in Neural Machine Translation Architectures



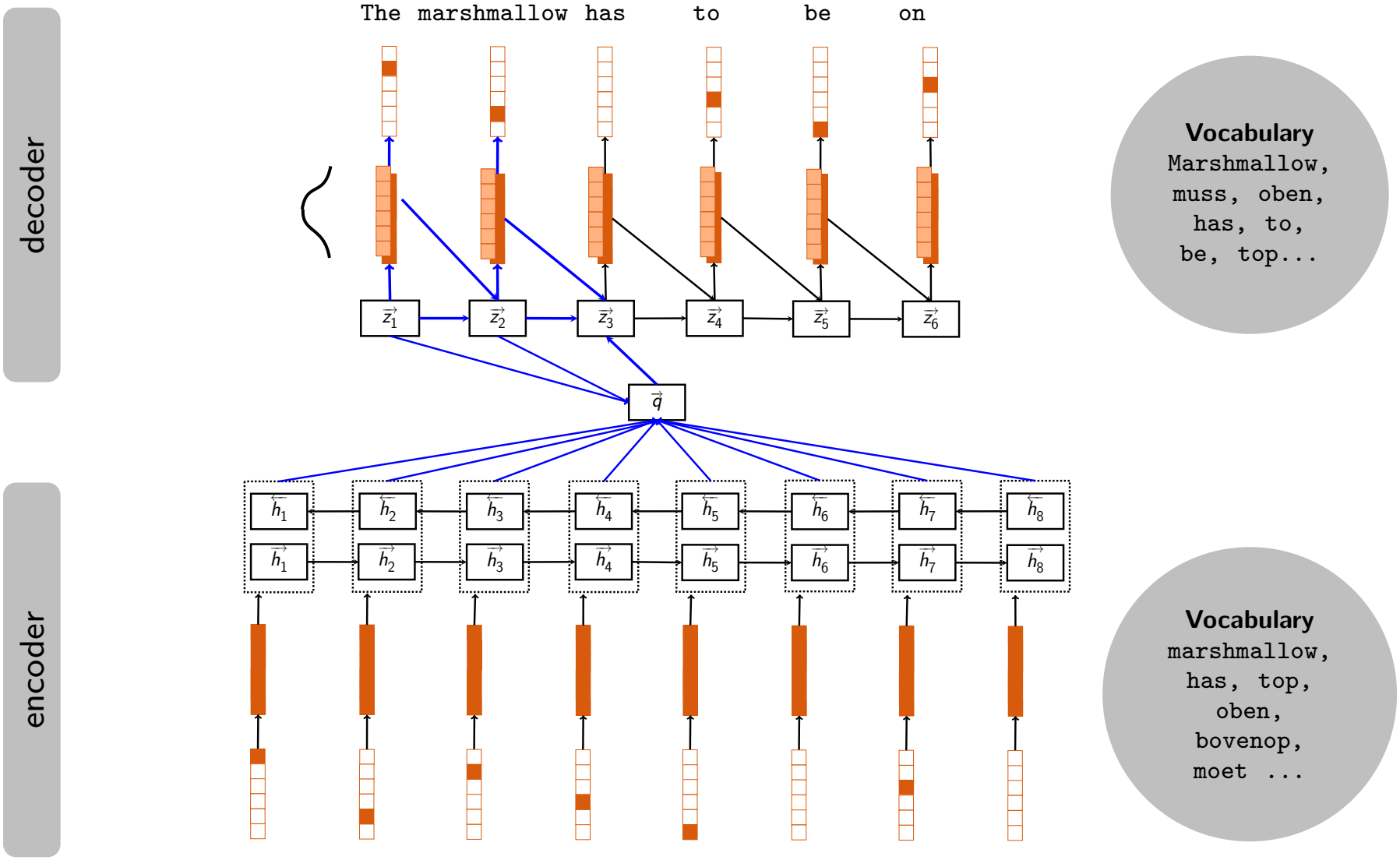
Semantic Representations

in Neural Machine Translation Architectures



Semantic Representations

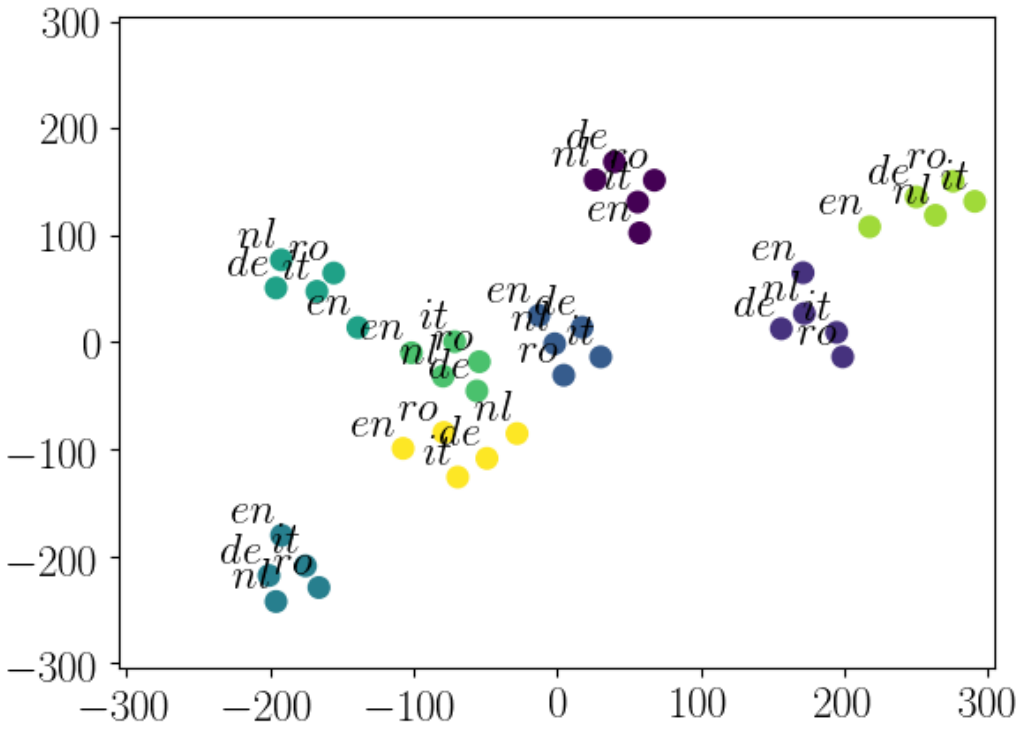
in Neural Machine Translation Architectures



Semantic Representations

Performance of Context Vectors

ML-NMT: $\{de, en, nl, it, ro\} \rightarrow \{de, en, nl, it, ro\}$

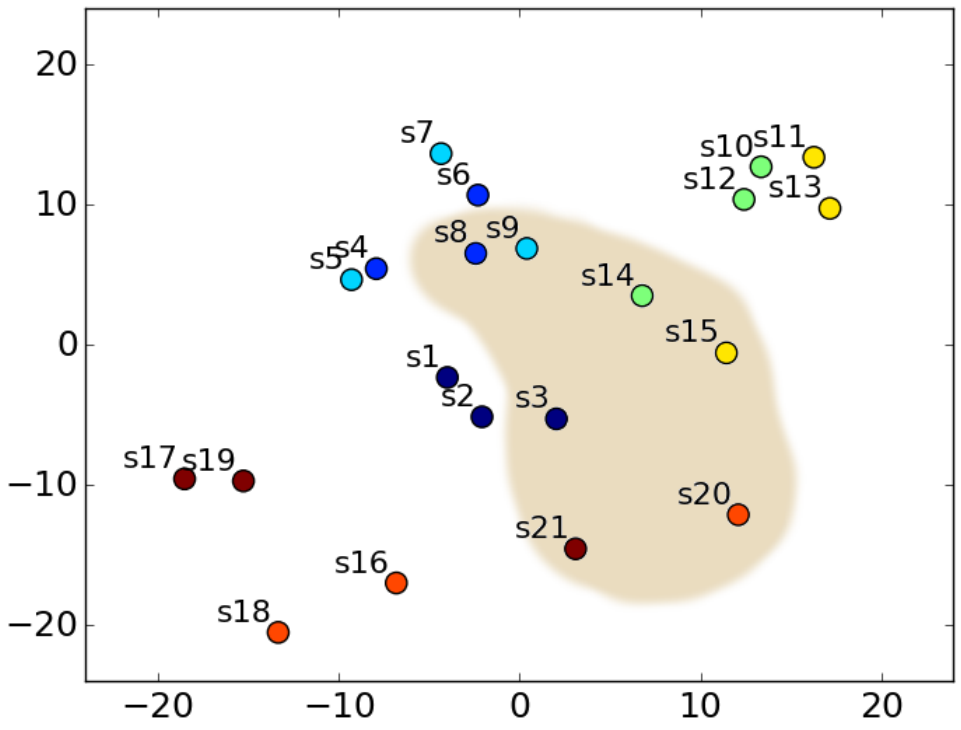


2D t-SNT representation of context vectors

Semantic Representations

Performance of Context Vectors

ML-NMT: $\{ar, en, es\} \rightarrow \{ar, en, es\}$



2D t-SNT representation of context vectors

Semantic Representations

Performance of Context Vectors

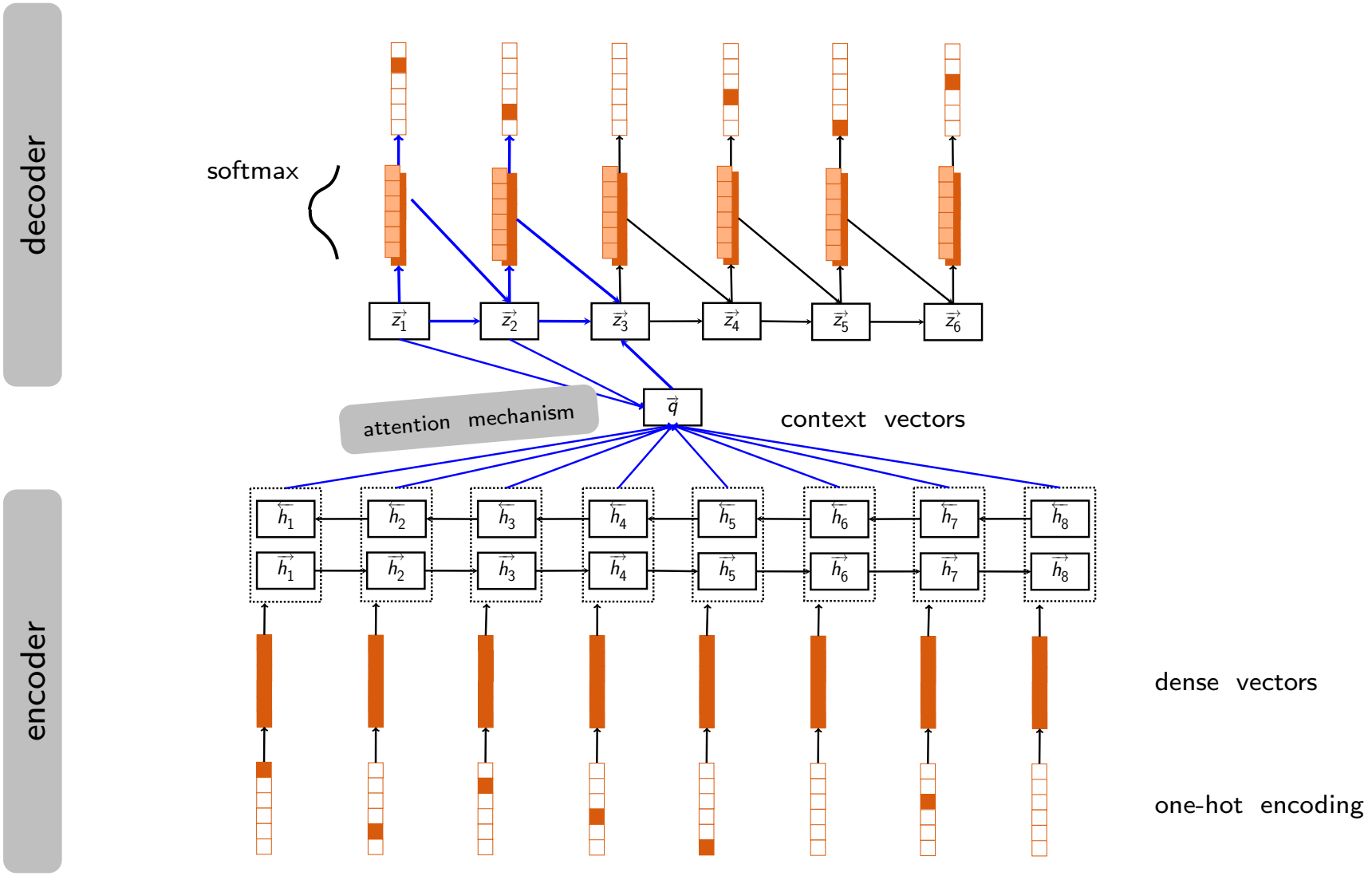
Semantic Textual Similarity Task (SemEval 2017)

	track1 *ar–ar*	track2 *ar–en*	track3 *es–es*	track4a *es–en*	track5 *en–en*
w2v 300-D	0.49	0.28	0.55	0.40	0.56
w2v 1024-D	0.51	0.33	0.59	0.45	0.60
ctx 1024-D	**0.59**	**0.44**	**0.78**	**0.49**	**0.76**

Pearson Correlation

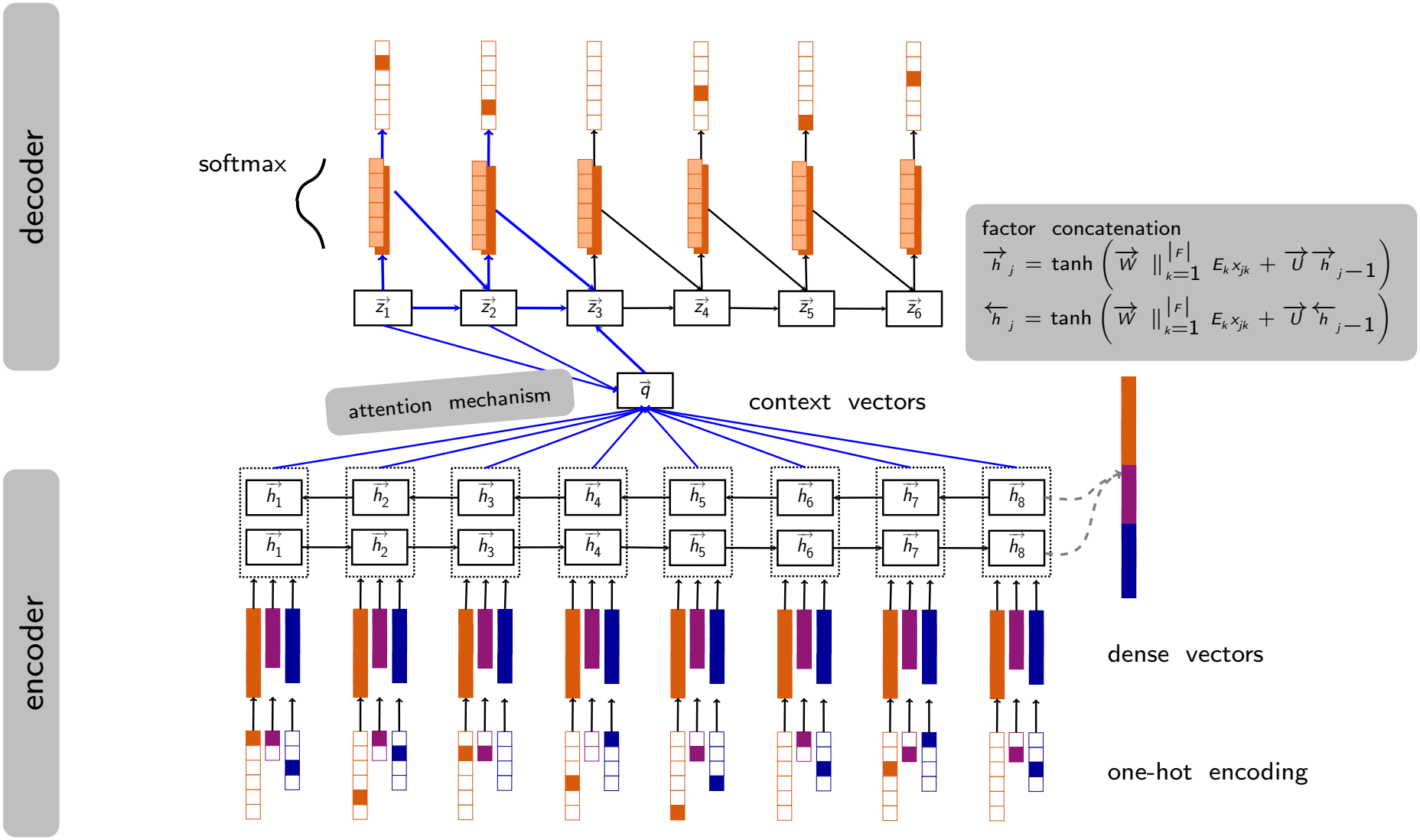
Semantic Representations

in Neural Machine Translation Architectures II



Semantic Representations

in Neural Machine Translation Architectures II



2de|-|- the|DET|O marsh|NOUN|MRXML has|VERB|HS to|PREP|T be|VERB|P on|PREP|AN top|NOUN|TP

Semantic Representations

Multilinguality and Interlinguality through Factors

- Approximate phonetic encodings with **Metaphone 3**
- **Babel Synsets** for nouns (incl. named entities, foreign words and numerals), adjectives, adverbs and verbs.
  Negation particles are tagged with `NEG`
- **PoS**, **Lemma** & **Stem**

Semantic Representations

Multilinguality and Interlinguality through Factors

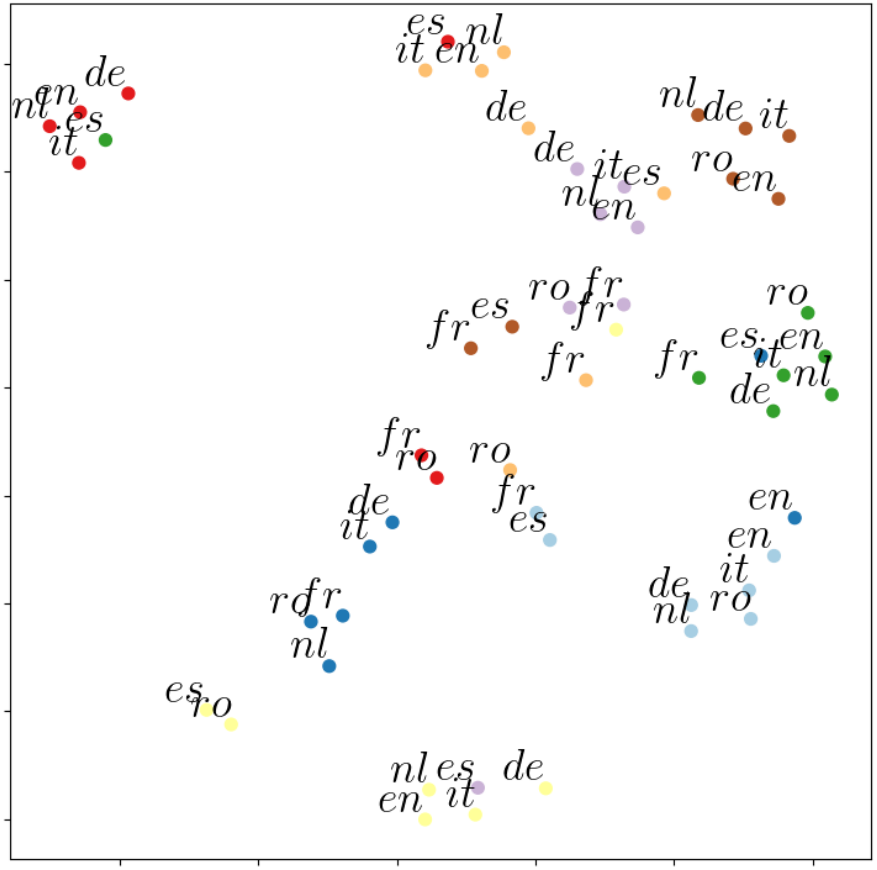
- Approximate phonetic encodings with **Metaphone 3**
- **Babel Synsets** for nouns (incl. named entities, foreign words and numerals), adjectives, adverbs and verbs.
  Negation particles are tagged with `NEG`
- **PoS**, **Lemma** & **Stem**

< 2*de* >|-|-|- the|DET|the|**0**|-
marshmallow|NOUN|marshmallow|**MRXML**|**bn:00053559n**
has|VERB|has|**HS**|**bn:00089240v** to|PREP|to|**T**|-
be|VERB|be|**P**|**bn:00083181v** on|PREP|on|**AN**|-
top|NOUN|top|**TP**|**bn:00077607n**

Semantic Representations

Beyond Zero-Shot Languages with Factors

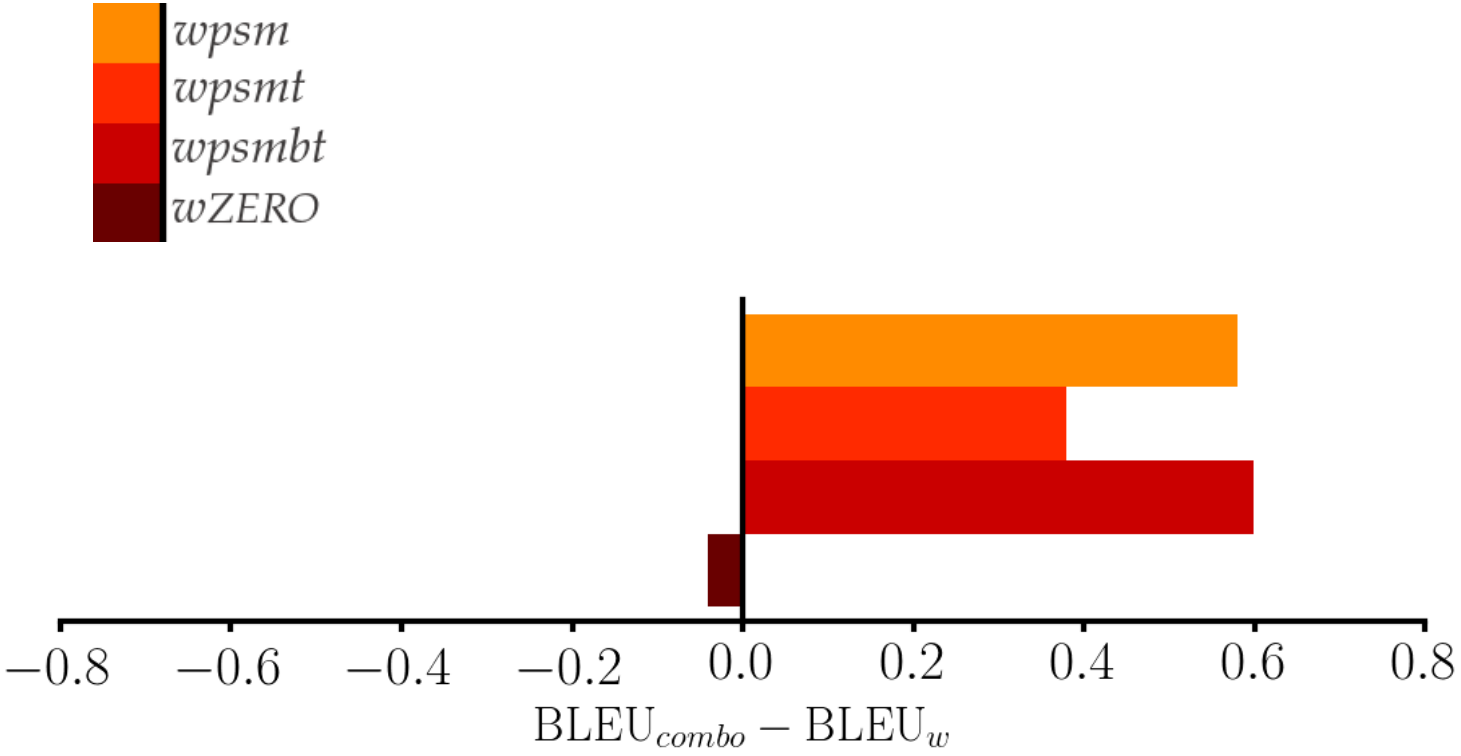
ML-NMT: $\{de, en, nl, it, ro\} \rightarrow \{de, en, nl, it, ro\} \not\rightarrow es, fr$!



2D t-SNT representation of context vectors

Semantic Representations

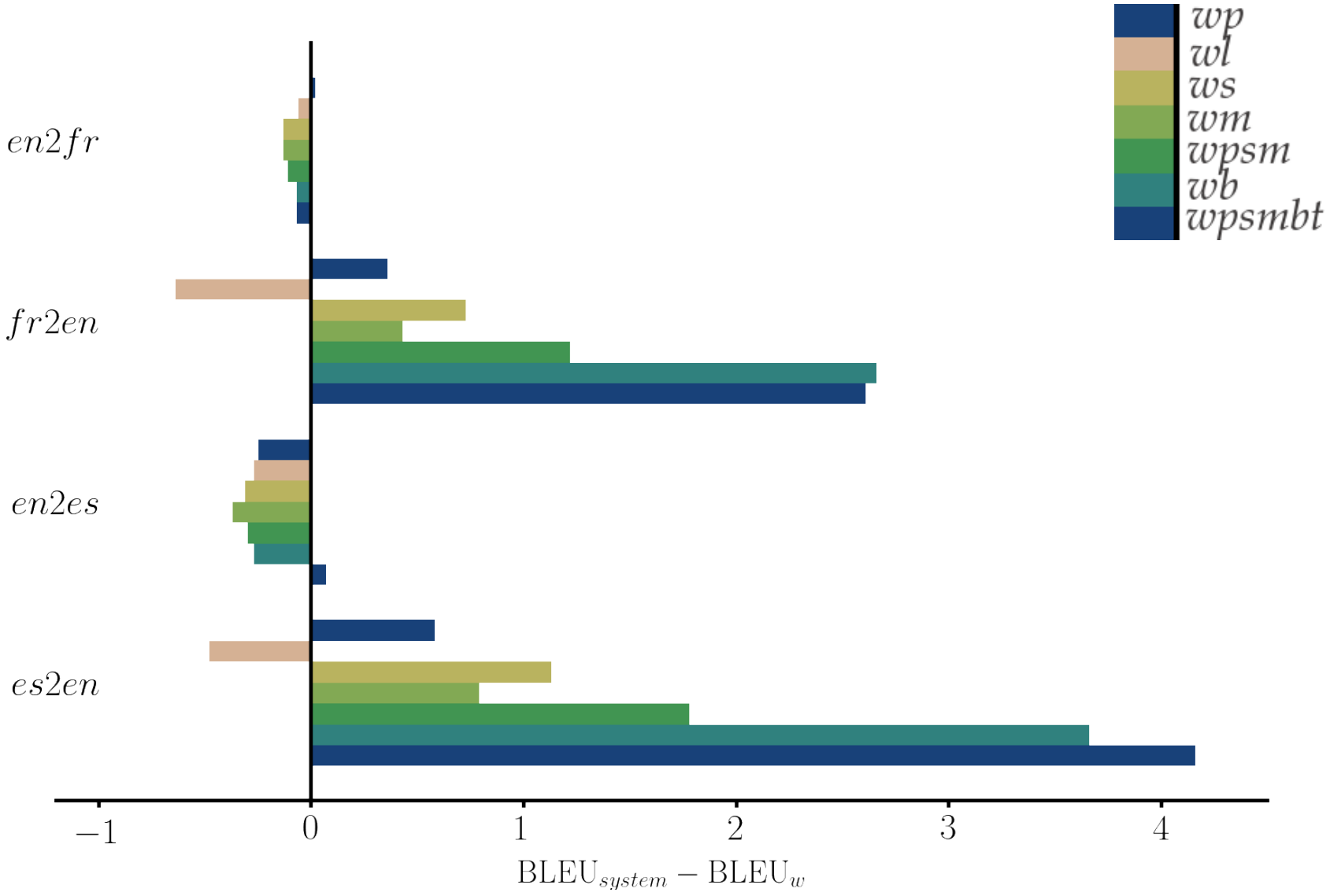
Improving NMT with Factors



Average for 20 language pairs

Semantic Representations

Beyond Zero-Shot NMT with Factors



Semantic Representations

...from Multilingual NMT Systems

- Context vectors provide a better **semantic representation at sentence level** than the composition of word embeddings
- They are multilingual by construction
- Interlingual factors help to position in the same space **new (related) languages**
- ... but you need **time**, **GPUs** and **data**

Multilingual Resources

Multilingual Resources

Importance of Resources

- Data driven systems are (even) more prevalent with the boom of **deep learning** architectures
- Data **quality** is (even) more important for neural systems e.g. SMT vs. NMT

Multilingual Resources

Multilingual Resources

BabelNet

ウィキペディア
フリー百科事典

Multilingual Resources

BabelNet

- Multilingual encyclopedic dictionary
- Semantic network
- 271 languages
- 14 million entries

Multilingual Word Sense Disambiguation

de: Es|- **war|bn:00083181v** ein|- riesiger|- Erfolg|bn:15350982n

en: And|- it|- **was|bn:00083181v** a|- huge|bn:00098905a success|bn:00075023n

it: Ed|- **è|bn:00083181v stato|bn:00083181v** un|- enorme|bn:00102268a successo|bn:00078365n

nl: En|- het|- **was|bn:00083181v** een|- groot|- succes|bn:06512571n

ro: Și|bn:00012706n a|- **fost|bn:00083181v** un|- mare|bn:00098342a succes|bn:00075024n

Multilingual Word Sense Disambiguation

de: Es|- **war|bn:00083181v** ein|- riesiger|- **Erfolg|bn:15350982n**

en: And|- it|- **was|bn:00083181v** a|- huge|bn:00098905a **success|bn:00075023n**

it: Ed|- **è|bn:00083181v stato|bn:00083181v** un|- enorme|bn:00102268a **successo|bn:00078365n**

nl: En|- het|- **was|bn:00083181v** een|- groot|- **succes|bn:06512571n**

ro: Şi|bn:00012706n a|- **fost|bn:00083181v** un|- mare|bn:00098342a **succes|bn:00075024n**

Multilingual Resources

Wikipedia

ウィキペディア
フリー百科事典

- Multilingual, web-based encyclopedia
- 299 languages
- 46 million articles
- *English*: 5,529,144 content articles;
  Japanese: 1,087,058 content articles

Multilingual Resources

Wikipedia: Related Research & Challenges

- Same article in multiple languages connected via **interlanguage links**
- **Comparable corpora** are easy to build

Multilingual Resources

Wikipedia: Related Research & Challenges

- Same article in multiple languages connected via **interlanguage links**
- **Comparable corpora** are easy to build
- Categories allow to extract **in-domain** comparable corpora (not so easy!)
- **Parallel sentences** can be extracted from comparable corpora (not so easy!)

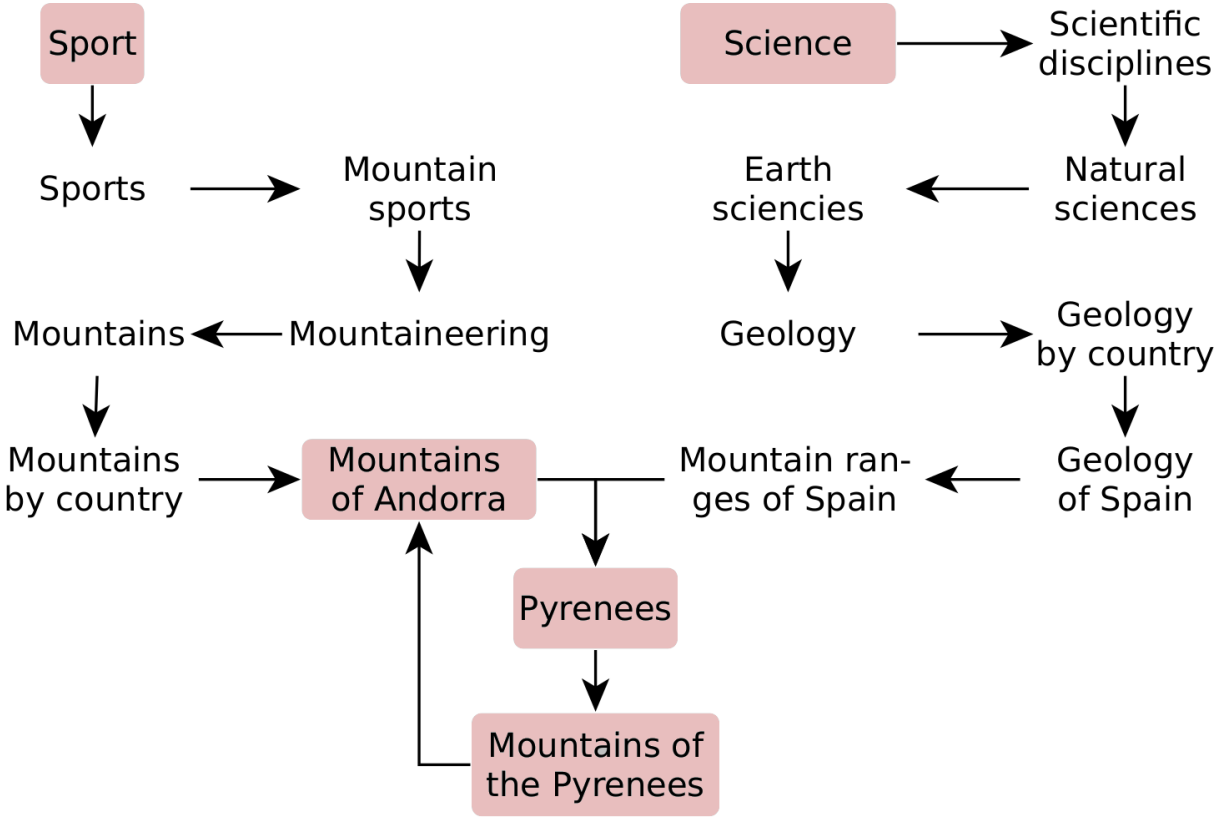
Multilingual Resources

Wikipedia: Related Research & Challenges

In-domain Comparable Corpora

- Identify comparable articles
- Build a characteristic vocabulary for the domain of interest automatically from root articles
- Explore the Wikipedia categories' graph to select the subset of categories in the domain

Graph Exploration



Multilingual Resources

Wikipedia: Related Research & Challenges

Parallel Sentence Extraction

- Brute-force sentence-wise comparison for parallel pairs identification
- Affordable (sentences aligned at document level)
- Similarity measures to detect them
  - Cosine on character *n*-grams and pseudo-cognates, length factors
  - Cosine on context vectors

Projects

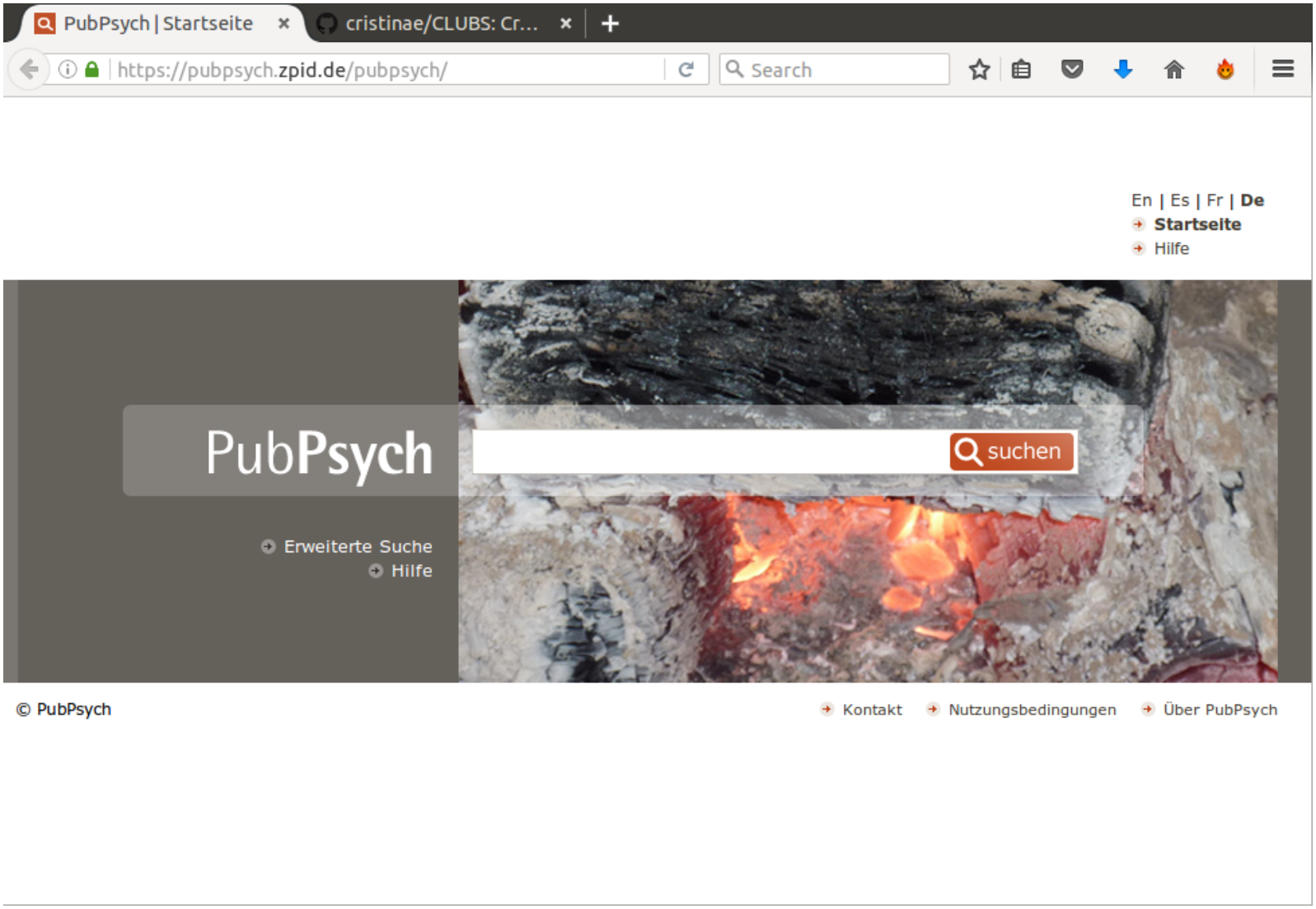
CLuBS

Cross-Lingual Bibliographic Search

- **Initial Status**: Multilingual *en*/*es*/*de*/*fr* retrieval platform —PubPsych— online
- **Aim**: Make PubPsych cross-lingual and improve translation and retrieval with the latest advances

Projects

PubPsych – `https://pubpsych.zpid.de/pubpsych/`



Projects

The CLuBS Project

- German Project (Leibniz Gemeinschaft)
  - Multilingual NMT systems
  - In-domain parallel sentence extraction
  - Query translation?
  - Cross-lingual retrieval?

Projects

The QT21 Project

- **Aim**: Improve translation quality for all European languages
- **Initial Status**: 3 well covered languages, problems with morphologically complex and low-resourced languages

Projects

The QT21 Project

- European Project (Horizon 2020)
  - Multilingual NMT systems
  - Factored ML-NMT with interlingual information

Projects

WikiTailor

WIKITAILOR

Your à-la-carte in-domain corpora extraction tool from Wikipedia

Joint work with Alberto Barrón-Cedeño

Projects

WikiTailor

WIKITAILOR

Your à-la-carte in-domain corpora extraction tool from Wikipedia

Joint work with Alberto Barrón-Cedeño

- **Aim**: Extraction of parallel corpora in any domain and language from Wikipedia
- **Current Status**: Extraction of comparable corpora in any domain and language from Wikipedia

Thanks!

Multilingual Natural Language Processing

Cristina España-Bonet

RICOH Institute of ICT, Tokyo, Japan

11th December 2017