Neural Machine Translation with Context & Document Information

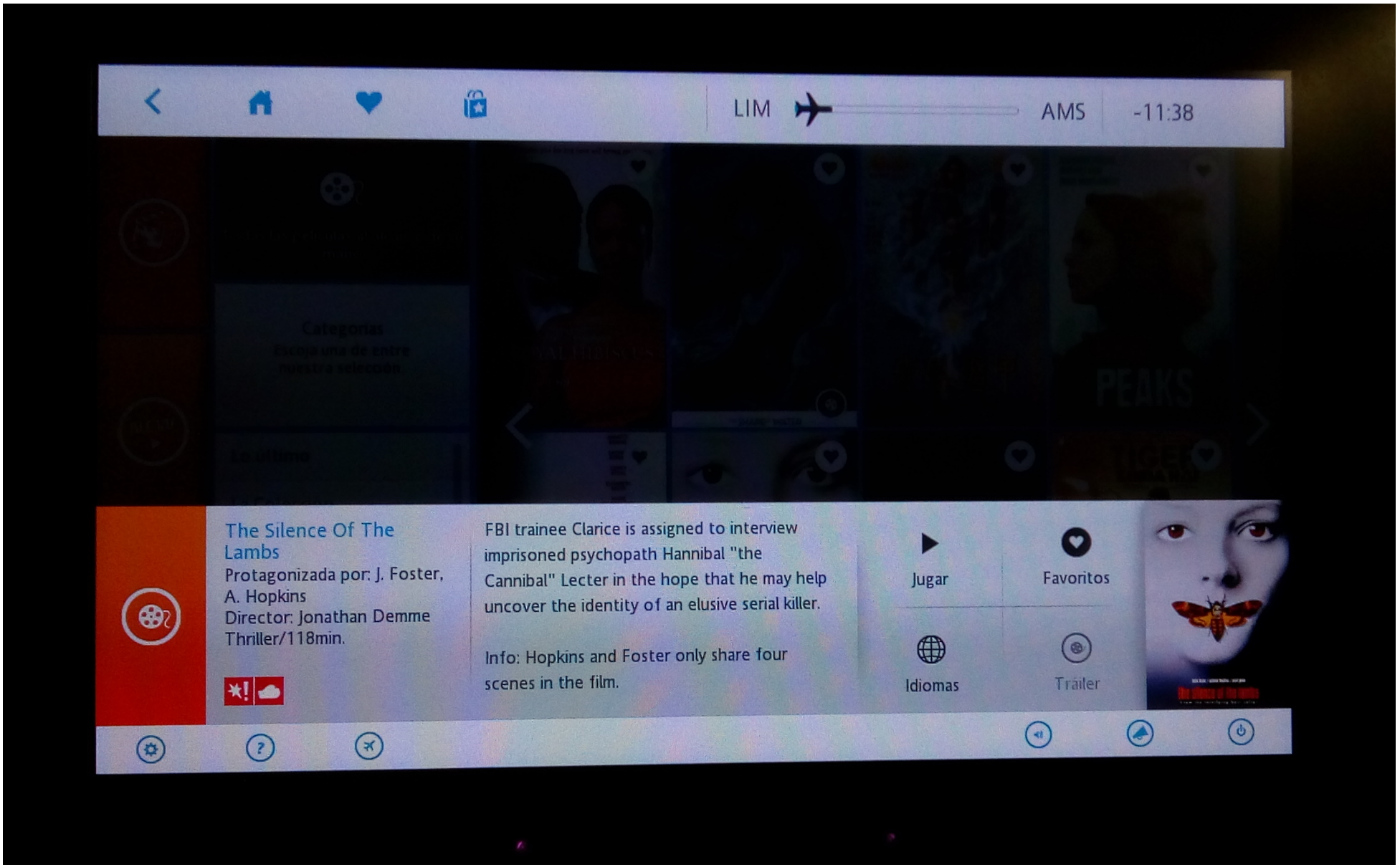
Cristina España-Bonet
Universität des Saarlandes (UdS)
Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI)

Guangdong University of Foreign Studies, China

23rd October 2018

Context vs. Document vs. Discourse MT

the screen in my flight



Context & Document NMT

Contents

Introduction

Motivation

MT systems usually operate at **sentence level** basis, so

1. Inter-sentence information is lost

   Consequence: e.g. bad coreference resolution

2. Document level information is lost

   Consequence: e.g. bad lexical selection

Introduction

Motivation II

Automatic translation

Li Na (tennis player)

In this Chinese name, the last name is Li .

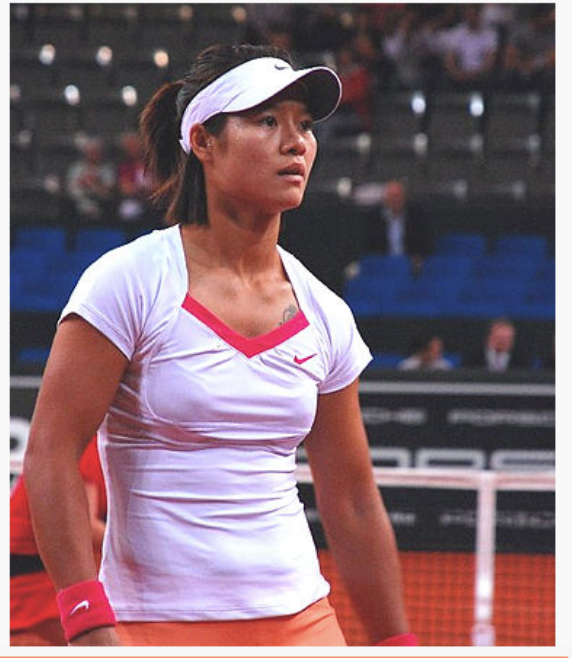
Li Na - Chinese : 李娜; pinyin : Lǐ Nà- (February 26, 1982 , Wuhan , China) is a Chinese professional tennis player who has won eight WTA titles in singles and two in doubles. His best position in the individual ranking was the fourth in 2011. After several years in the women's circuit with quite mediocre results, he was exposed to a fairly advanced age (28) by professional tennis players, so that 2011 was finalist at the Austrian Open and winner at Roland Garros . With this milestone he became the first Asian tennis player to win an individual Grand Slam title. He also won the Australian Open 2014 .

Content [amaga]

1 Origins

2 Grand Slam tournaments

Li Na



Introduction

Motivation III

Source

Li Na (tennista)

En aquest nom xinès el cognom és Li.

Li Na —xinès: 李娜; pinyin: Lǐ Nà— (26 de febrer del 1982, Wuhan, Xina) és una jugadora de tennis professional xinesa que ha guanyat vuit títols WTA en individuals i dos en dobles. La seva millor posició en el rànquing individual fou la quarta l'any 2011. Després de diversos anys en el circuit femení amb resultats força mediocres, es va destapar a una edat força avançada (28) pels tennistes professionals, de manera que el 2011 fou finalista a l'Open d'Austràlia i vencedora al Roland Garros. Amb aquesta fita esdevingué la primera tennista asiàtica en guanyar un títol individual de Grand Slam. Posteriorment va guanyar també l'Open d'Austràlia 2014.

Contingut [amaga]

1 Orígens

Li Na

Introduction

Proposal

Translate at sentence level but

1. Detect and enrich the system with coreference chains along a document

   Tackles wrong coreference resolution

2. Enrich sentences with information on the topic of the document they belong to

   Tackles wrong lexical selection

Introduction

Big Pros & Cons

✓ Sentence-level decoding still possible

✗ Parallel corpora at document level needed in order to train a NMT system

Introduction

Going Besides Sentence Level

Architecture-based methods:

- **Exploiting cross-sentence context for neural machine translation**
  Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu.
  pages 2826–2831, EMNLP 2017

- **Topic-informed neural machine translation**
  Jian Zhang, Liangyou Li, Andy Way, and Qun Liu.
  pages 1807–1817, COLING 2016

Introduction

Going Besides Sentence Level

Architecture-based methods:

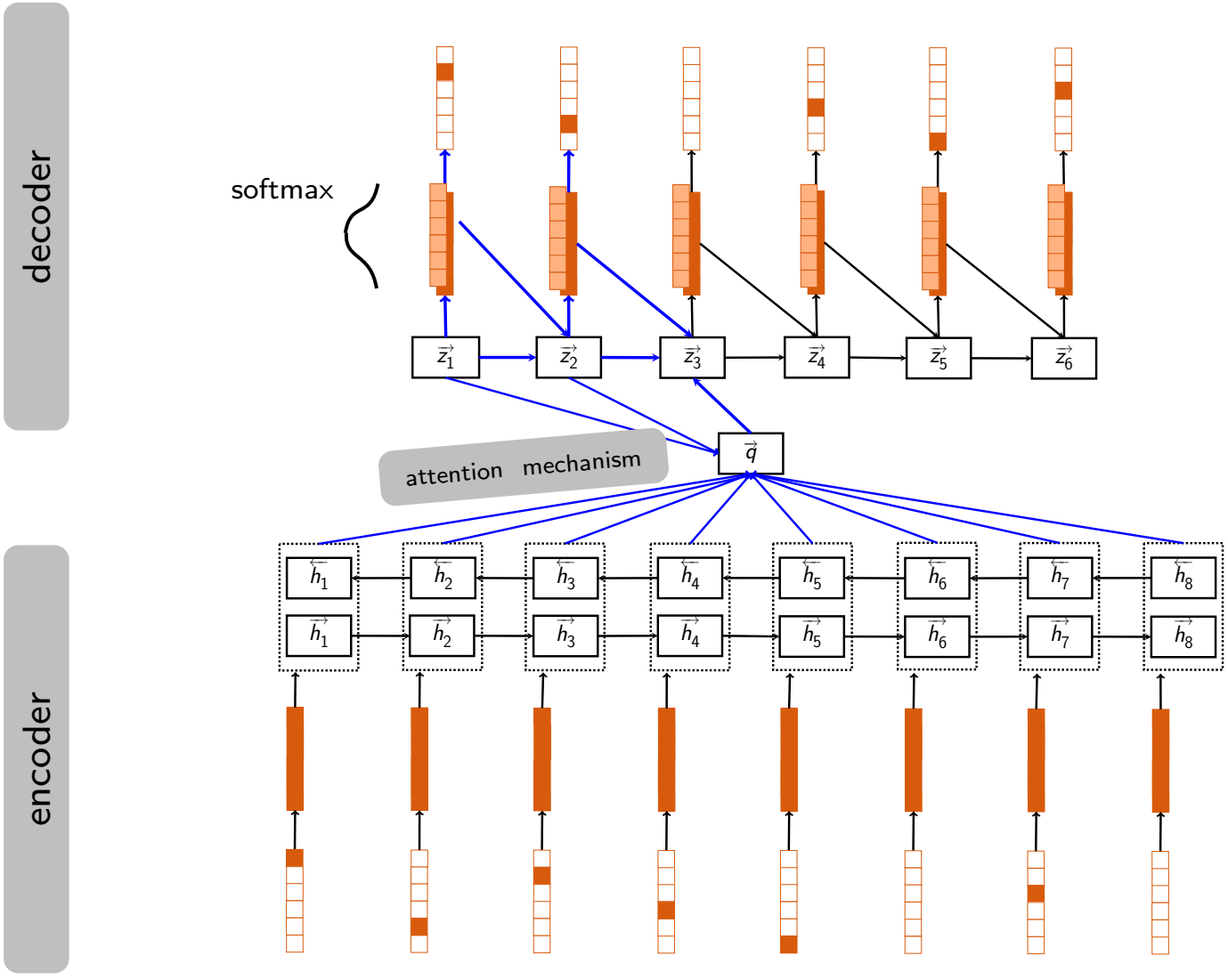
- **Exploiting cross-sentence context for neural machine translation**
  Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu.
  pages 2826–2831, EMNLP 2017

- **Topic-informed neural machine translation**
  Jian Zhang, Liangyou Li, Andy Way, and Qun Liu.
  pages 1807–1817, COLING 2016

Data-based methods...

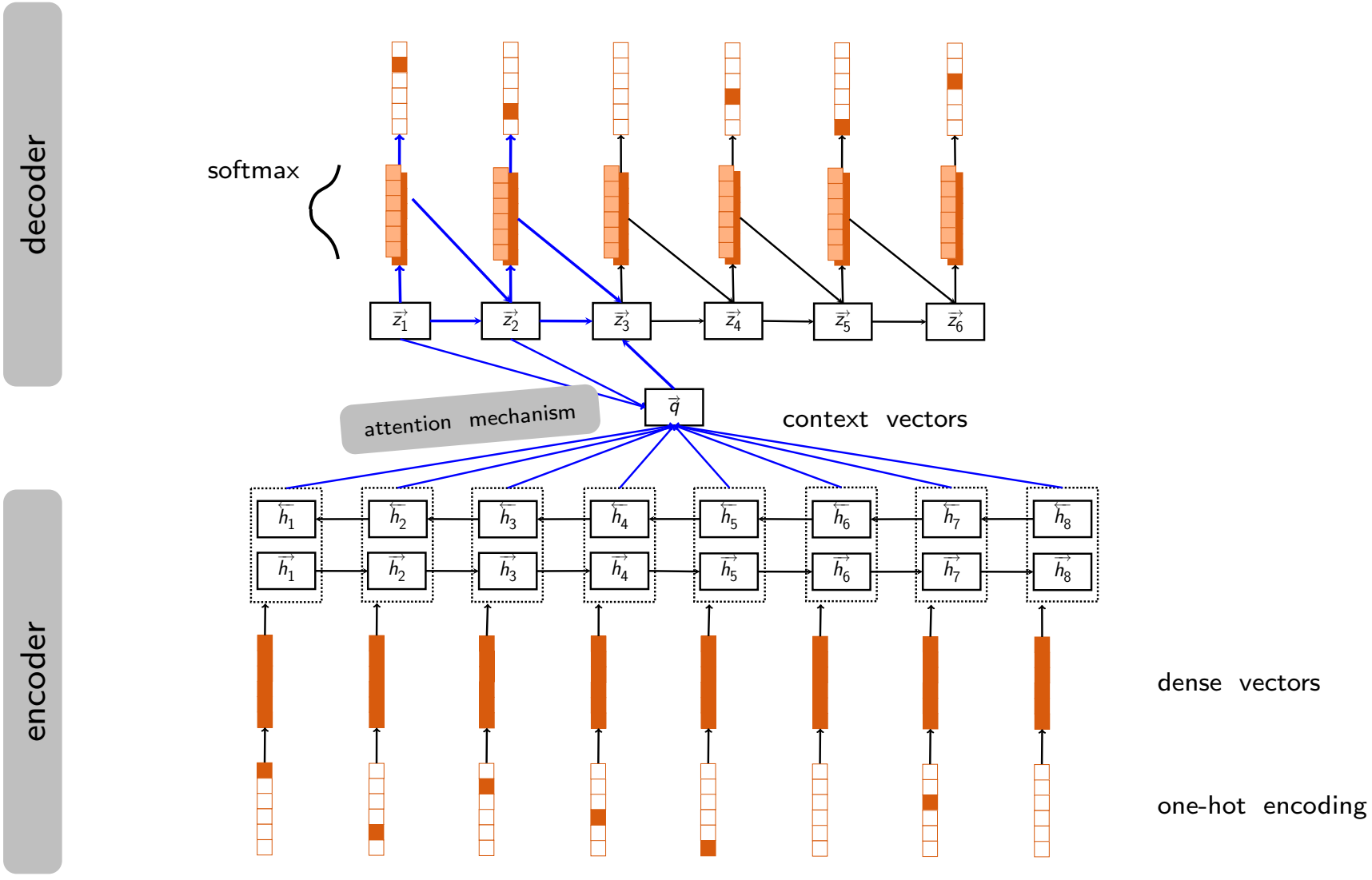
Neural Machine Translation

Adding Additional Information via Tags



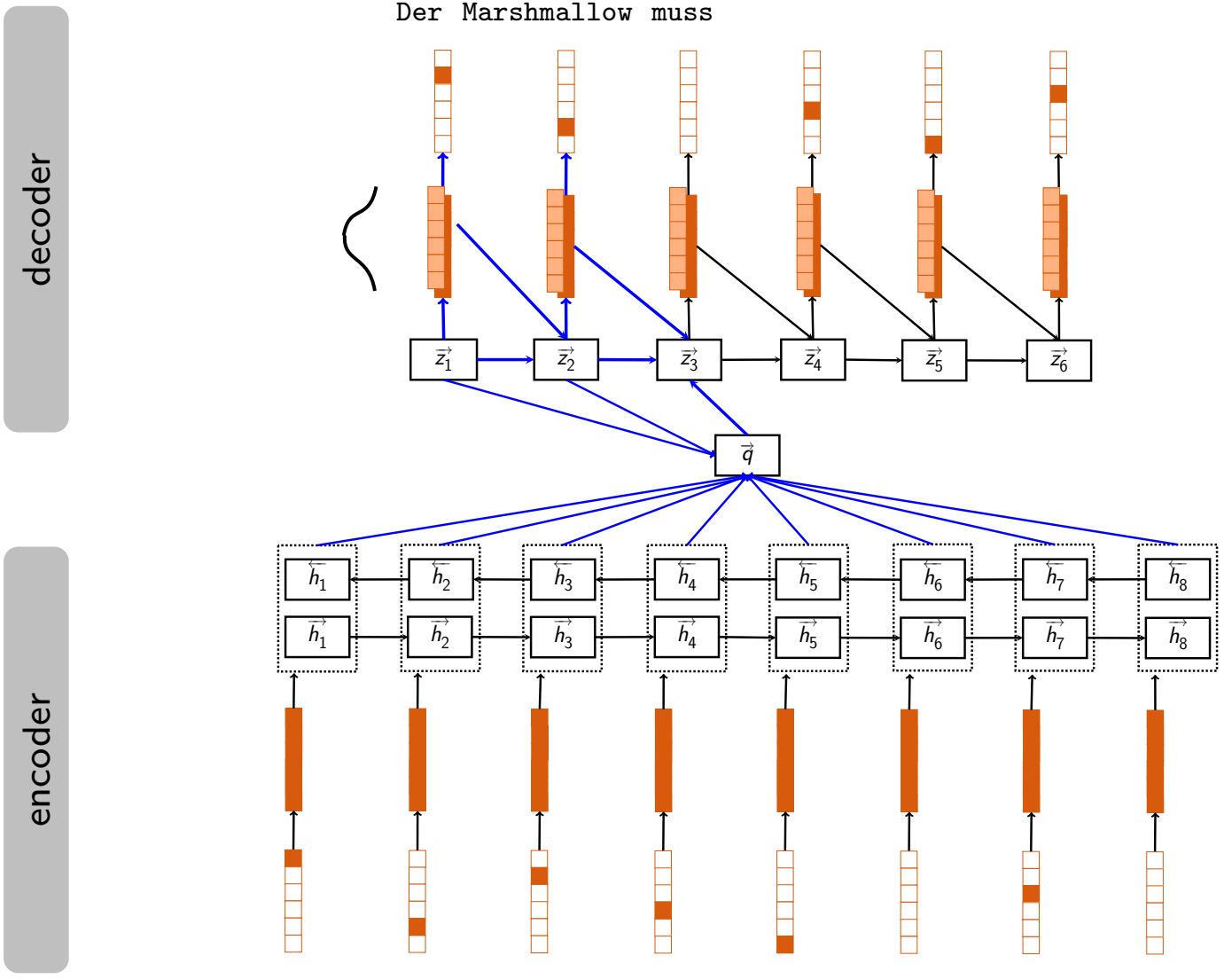
Neural Machine Translation

Adding Additional Information via Tags



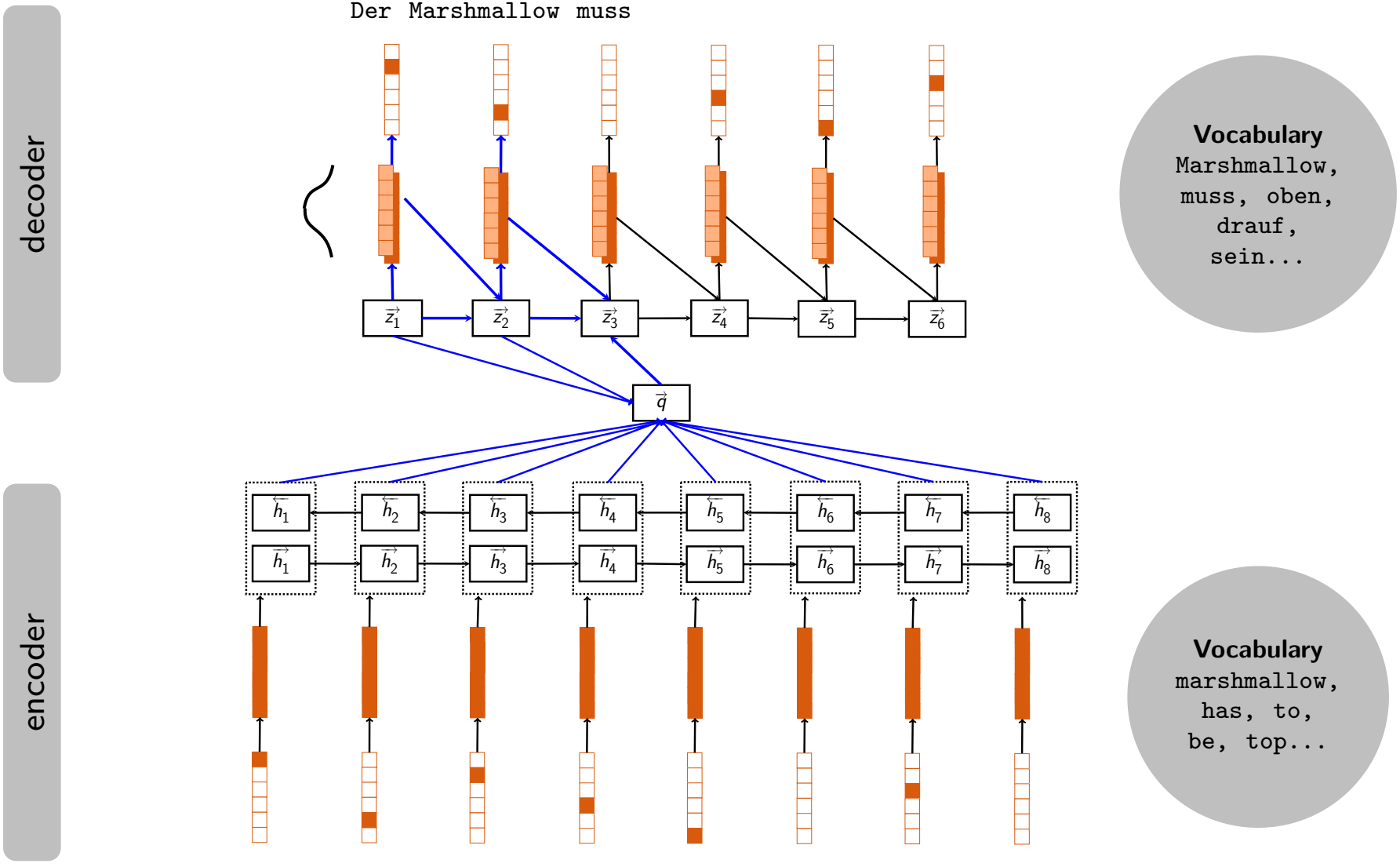
Neural Machine Translation

Adding Additional Information via Tags



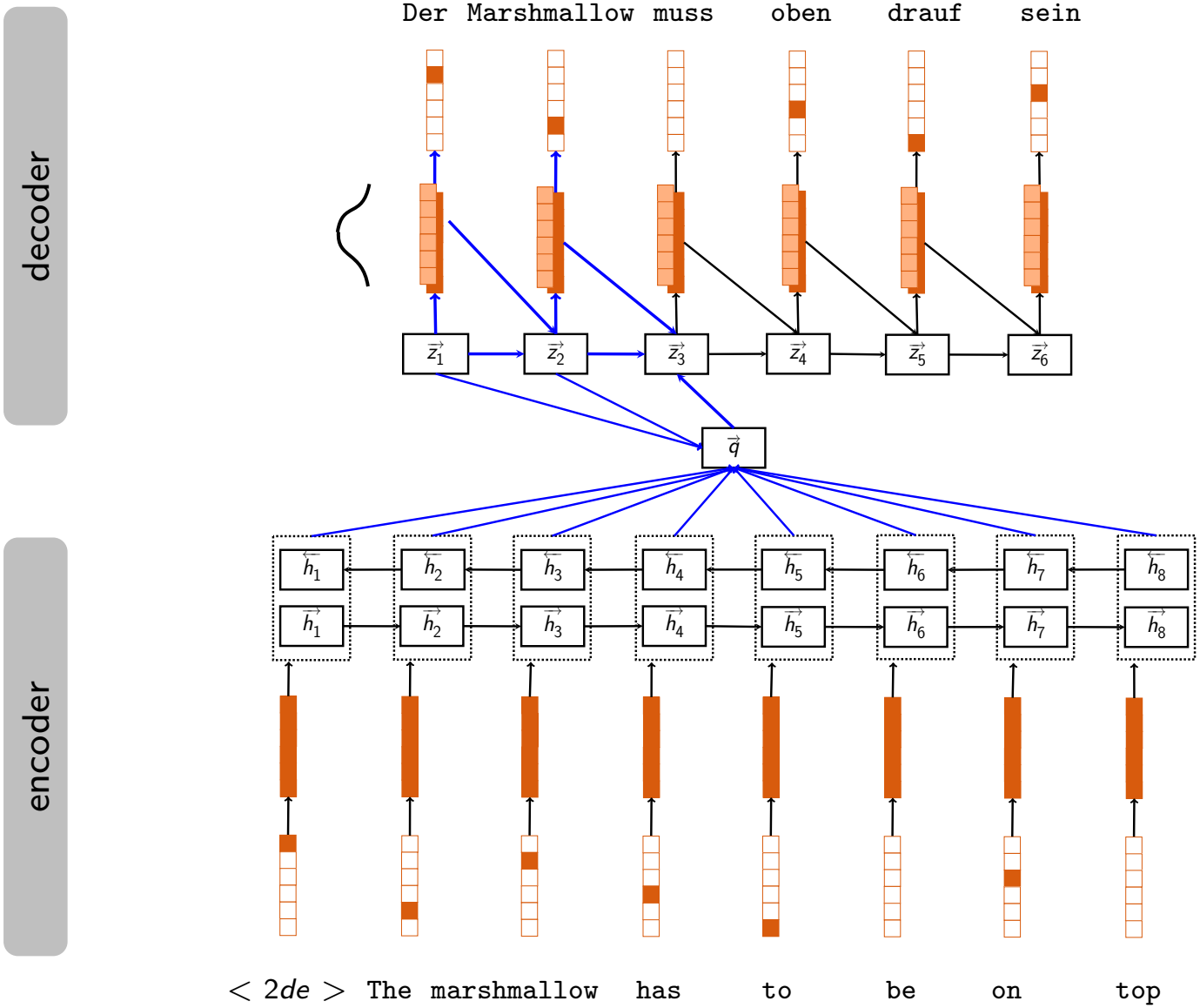
Neural Machine Translation

Adding Additional Information via Tags



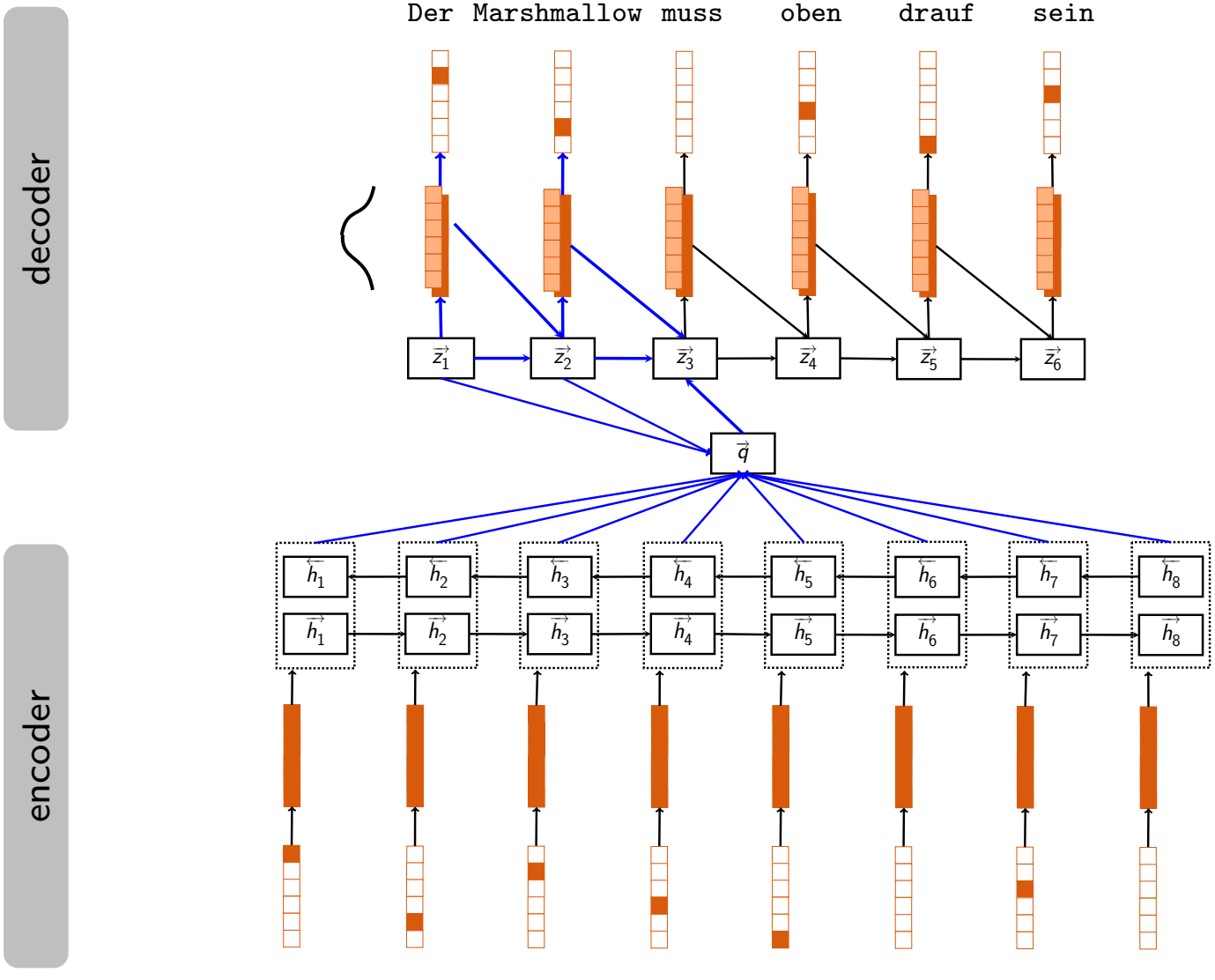
Neural Machine Translation

Adding Additional Information via Tags



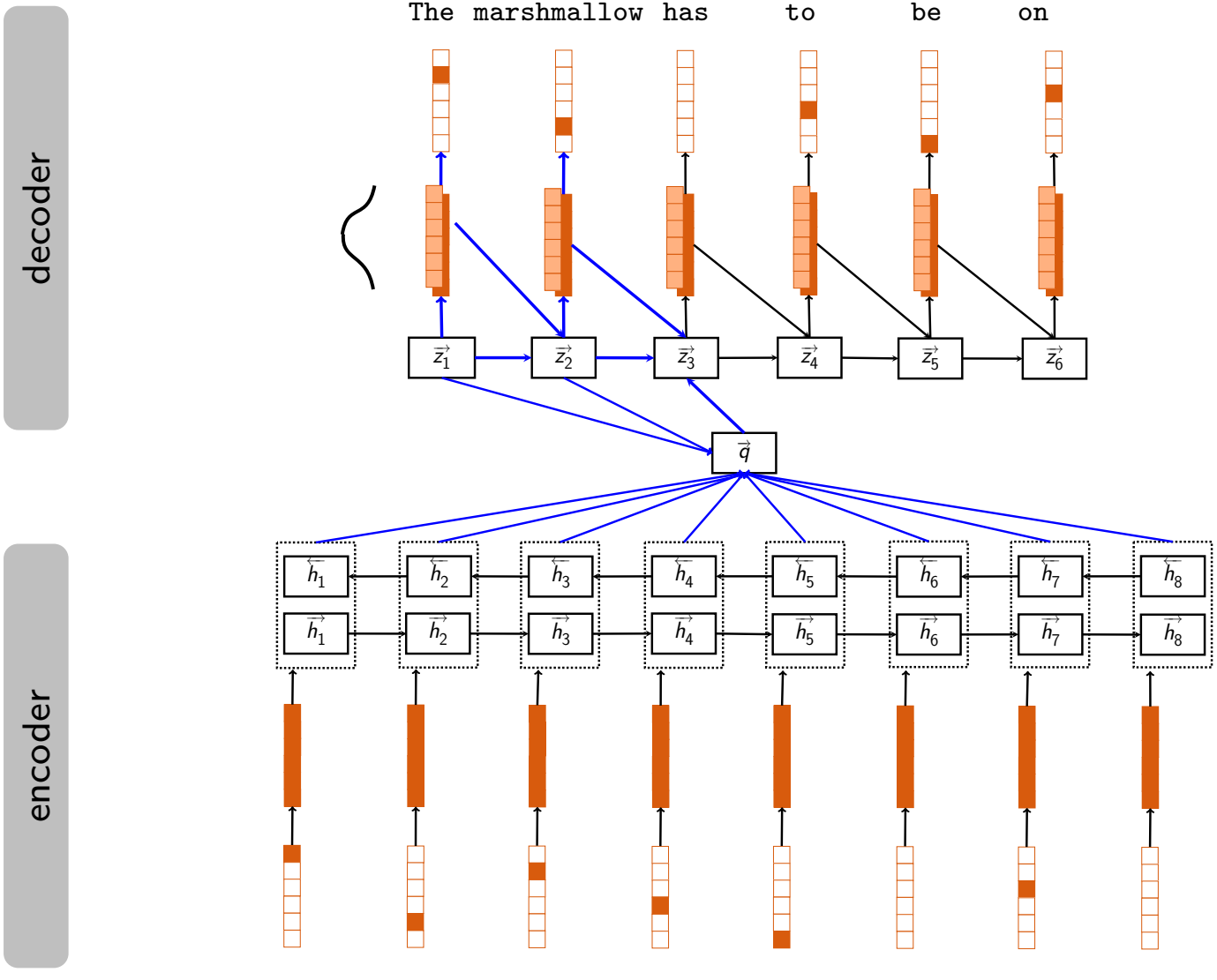
Neural Machine Translation

Adding Additional Information via Tags



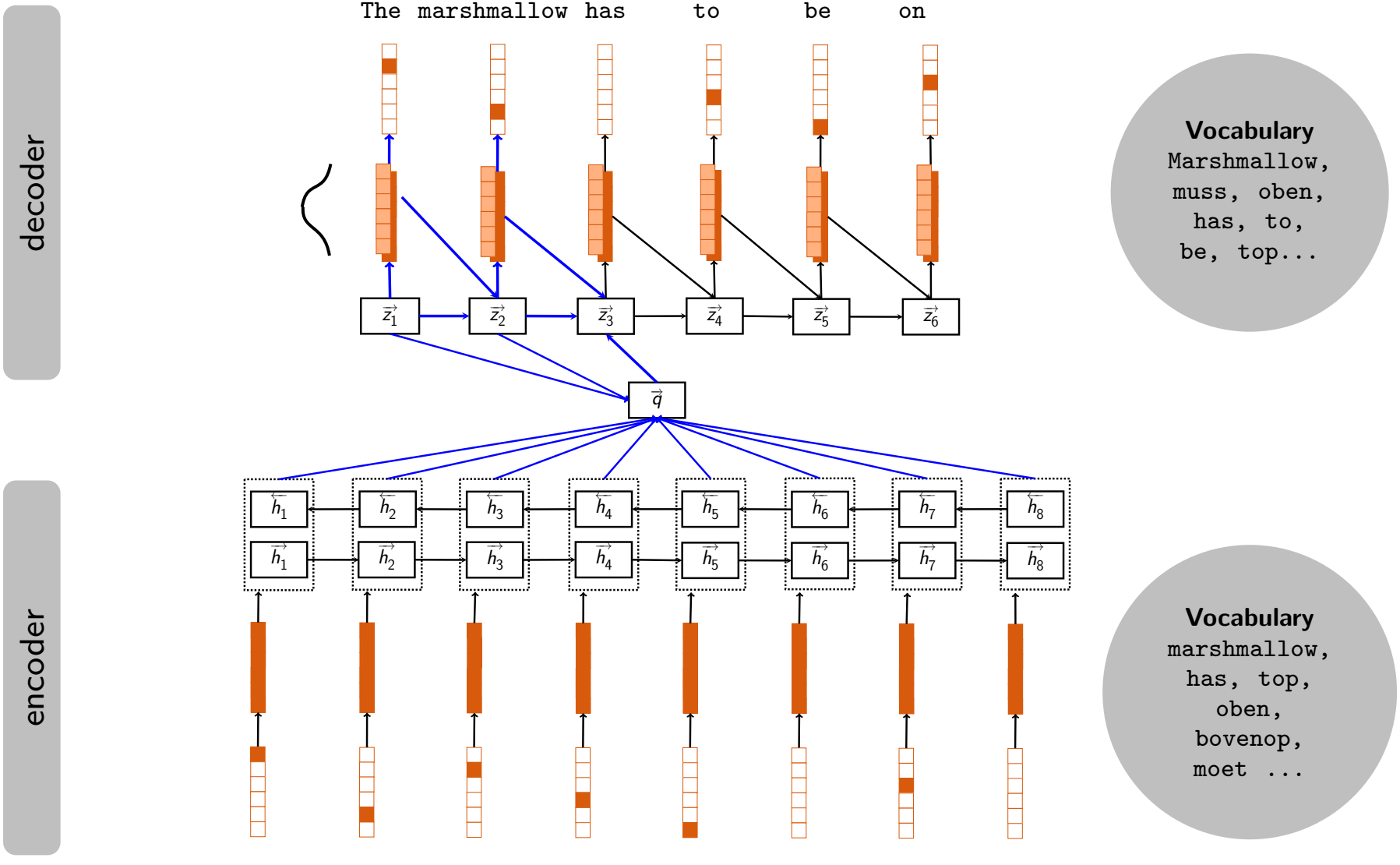
Neural Machine Translation

Adding Additional Information via Tags



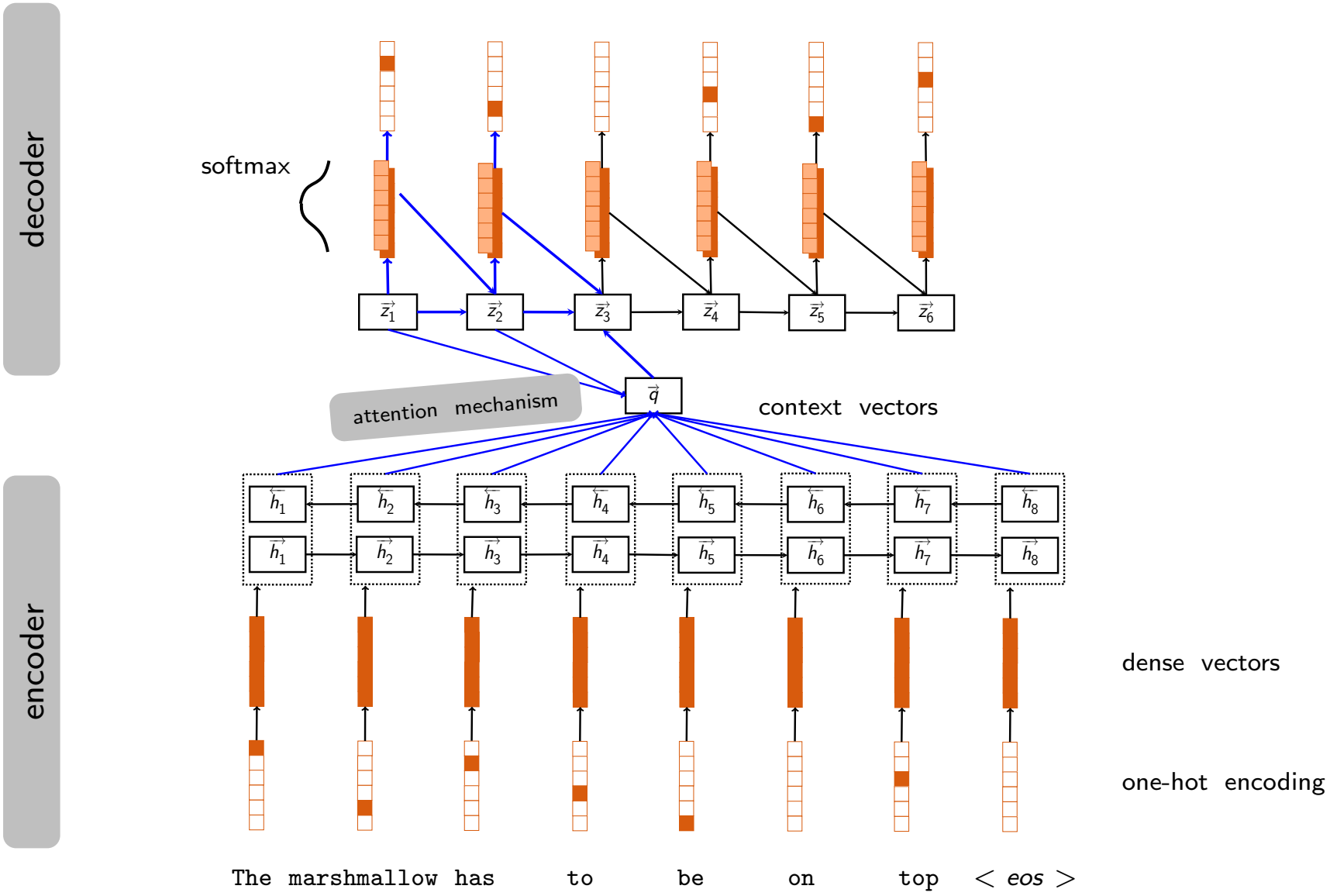
Neural Machine Translation

Adding Additional Information via Tags



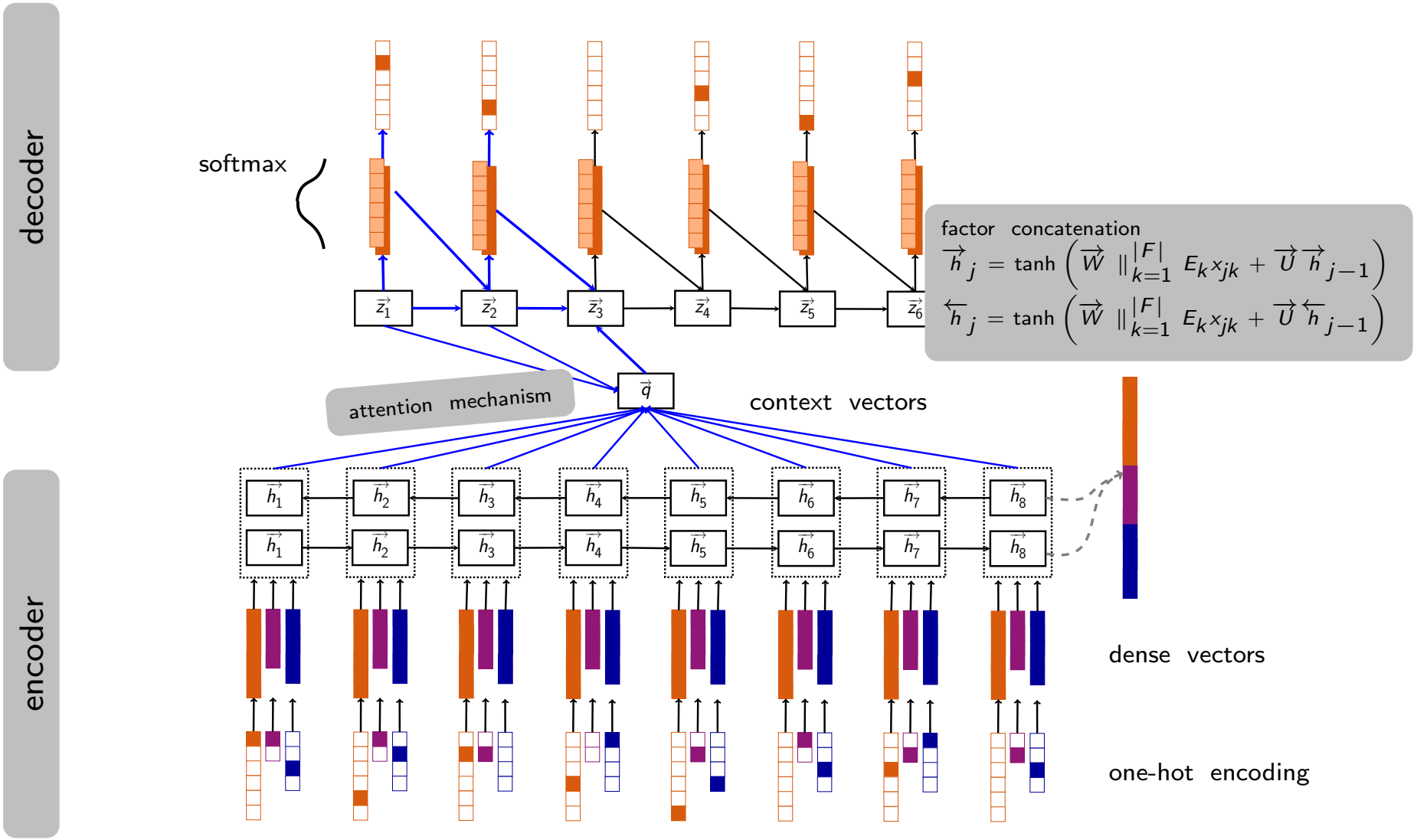
Neural Machine Translation

Adding Additional Information via Factors



Neural Machine Translation

Adding Additional Information via Factors



2de|-|- the|DET|0 marsh|NOUN|MRXML has|VERB|HS to|PREP|T be|VERB|P on|PREP|AN top|NOUN|TP

Neural MT Systems

Contents

Data

Corpora & Settings

Main figures

- en2es & en2de experiments
- Training: 25,269 documents en2de and 26,246 en2es
  ($\sim$ 4 M parallel sentences)
- Testing: newstest2013 (52 docs), TEDtest2010 (11 docs)

Data

Corpora & Settings II

Software

- **Nematus** for NMT training and decoding
- **Stanford CoreNLP** for coreference resolution
  - Neural-network-based mention-ranking model
  - P=80% and R=70% on the CoNLL 2012 English Test Data
- **Mallet** for topic modelling
  - Latent Dirichlet Allocation algorithm

Data

Corpora collection

English–Spanish

Corpus	Docs.	Sentences	Mentions	Used
Europarl	3,873	1,998,678	4,675,865	1,148,313
NewsCommentary	6,845	258,101	429,156	140,449
TED	1,795	219,393	451,724	114,826
JRC-Acquis	11,840	805,451	1,049,983	347,293
EMEA	1,893	870,403	352,051	110,854

Data

Corpora collection

English–German

Corpus	Docs.	Sentences	Mentions	Used
Europarl	3,622	1,952,708	4,570,736	1,123,031
NewsCommentary	6,257	240,726	389,878	127,596
TED	1,590	198,290	398,414	101,022
JRC-Acquis	11,889	719,080	1,046,329	350,224
EMEA	1,910	842,962	335,594	105,832

Experiments

Translation in Context: Coreferences I

- Source side (English) annotated with coreferences
- Mentions are detected and annotated with the head of the chain
  - We only consider mentions with less than 4 words,
  - *the* and *'s* are removed from the head if present,
  - we do not enrich the pronoun *I*

Experiments

Translation in Context: Coreferences I

- Source side (English) annotated with coreferences
- Mentions are detected and annotated with the head of the chain
  - We only consider mentions with less than 4 words,
  - *the* and *'s* are removed from the head if present,
  - we do not enrich the pronoun *I*

Example

baseline: I almost never cook with it.

coref: I almost never cook with <*b_crf*> *fish skin* <*e_crf*> it.

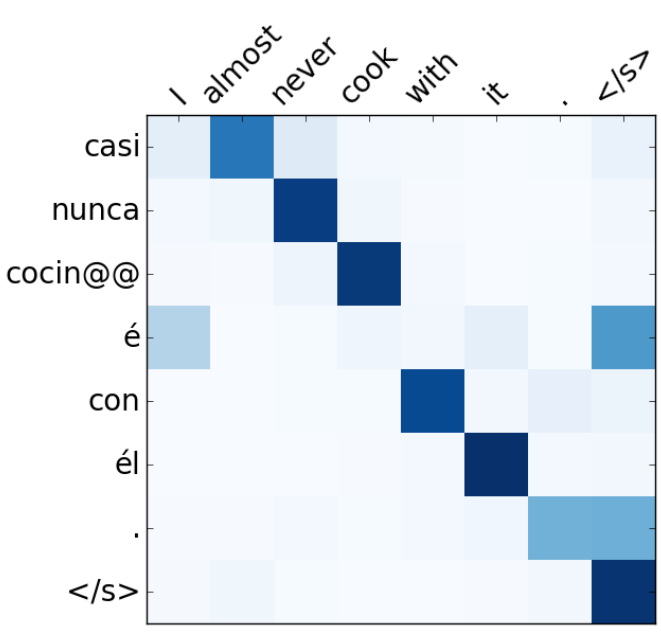
Experiments

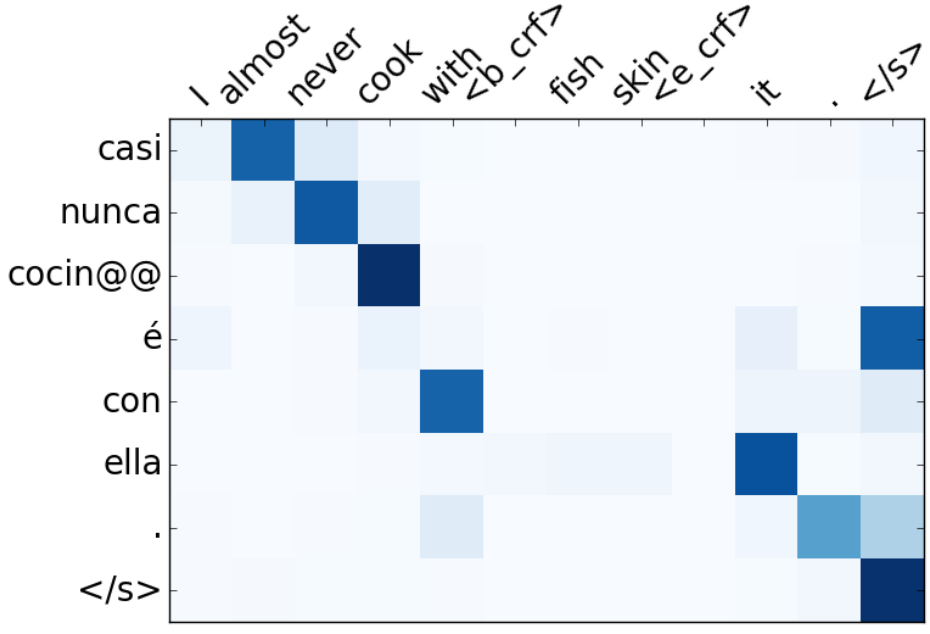
Translation in Context: Coreferences II

	En–De			En–Es		
	TER	BLEU	MTR	TER	BLEU	MTR
NEWS *baseline*	**60.27**	22.04	40.73	50.65	29.93	52.10
NEWS *coref*	60.37	**22.26**	**40.84**	**50.54**	**30.09**	**52.34**
TED *baseline*	**51.78**	**27.40**	46.00	41.21	37.44	59.48
TED *coref*	51.89	27.17	**46.21**	**41.05**	**37.88**	**59.75**

Experiments

Translation in Context: Coreferences III





Experiments

Translation in Context: Coreferences IV

Source:

However, it was later revised.

Experiments

Translation in Context: Coreferences IV

Source:

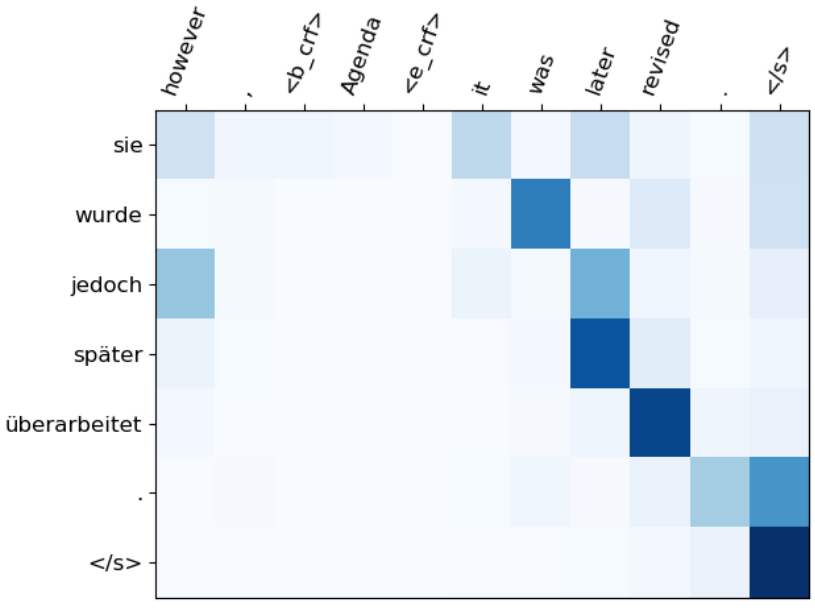
However, it was later revised.

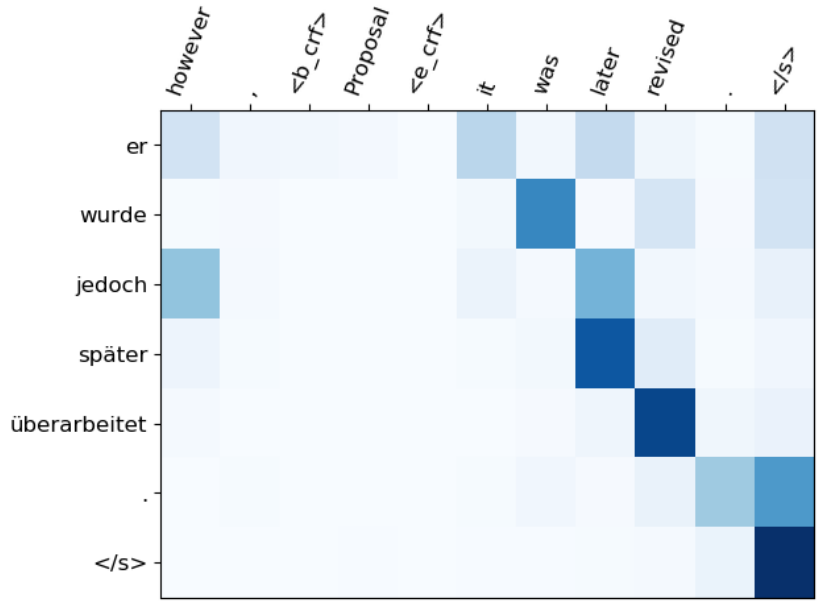
Translation in context:

- The agenda was adopted.
  However, <*b_crf*> *agenda* <*e_crf*> it was later revised.
- The proposal was adopted.
  However, <*b_crf*> *proposal* <*e_crf*> it was later revised.

Experiments

Translation in Context: Coreferences IV



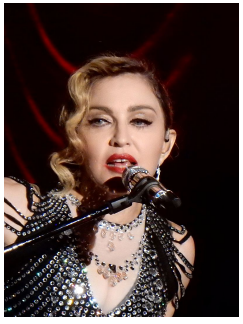


Sie wurde jedoch später überarbeitet.

Er wurde jedoch später überarbeitet.

Experiments

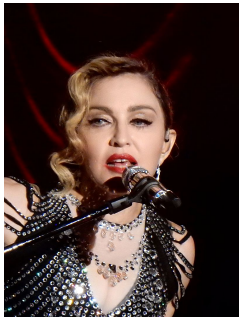
Translation in Context: Coreferences V



Madonna Louise Ciccone is an American singer, songwriter, actress, and businesswoman. Referred to as the "Queen of Pop" since the 1980s, *<b_crf> Madonna Louise Ciccone <e_crf>* **Madonna** is known for pushing the boundaries of songwriting in mainstream popular music, as well as imagery in music videos and on stage. *<b_crf> Madonna Louise Ciccone <e_crf>* **She** has also frequently reinvented both *<b_crf> Madonna Louise Ciccone <e_crf>* **her** music and image while maintaining autonomy within the recording industry. Besides sparking controversy, *<b_crf> Madonna Louise Ciccone <e_crf>* **her** works have been praised by music critics. *<b_crf> Madonna Louise Ciccone <e_crf>* **Madonna** is often cited as an influence by other artists .

Experiments

Translation in Context: Coreferences V



Madonna Louise Ciccone is an **American singer**, **songwriter**, **actress**, and **businesswoman**. Referred to as the "Queen of Pop" since the 1980s, *<b_crf> Madonna Louise Ciccone <e_crf>* **Madonna** is known for pushing the boundaries of songwriting in mainstream popular music, as well as imagery in music videos and on stage. *<b_crf> Madonna Louise Ciccone <e_crf>* **She** has also frequently reinvented both *<b_crf> Madonna Louise Ciccone <e_crf>* **her** music and image while maintaining autonomy within the recording industry. Besides sparking controversy, *<b_crf> Madonna Louise Ciccone <e_crf>* **her** works have been praised by music critics. *<b_crf> Madonna Louise Ciccone <e_crf>* **Madonna** is often cited as an influence by other artists .

Experiments

Translation with Document Level Information: Topics I

- Source side (English) annotated with topics
- Each word/sentence is annotated with the topic of the document it belongs to
  - Model with 7 topics
  - Model with 100 topics

Experiments

Translation with Document Level Information: Topics I

- Source side (English) annotated with topics
- Each word/sentence is annotated with the topic of the document it belongs to
  - Model with 7 topics
  - Model with 100 topics

Example

baseline: I almost never cook with it.
topic7: $< 5\_t >$ I almost never cook with it.
topic100: $< 68\_t >$ I almost never cook with it.
topic7fact: I|5_*t* almost|5_*t* never|5_*t* cook|5_*t* with|5_*t* it|5_*t*

Experiments

Translation with Document Level Information: Topics II

	En–De		
	TER	BLEU	MTR
NEWS *baseline*	**60.27**	**22.04**	**40.73**
NEWS *topic7*	61.06	21.91	40.48
NEWS *topic7fact*	60.78	21.72	40.52
NEWS *topic100*	61.30	21.59	40.43
NEWS *topic100fact*	60.88	21.98	40.66
TED *baseline*	51.78	27.40	46.00
TED *topic7*	51.69	27.24	46.12
TED *topic7fact*	**51.43**	**27.79**	**46.46**
TED *topic100*	52.26	26.98	45.91
TED *topic100fact*	52.32	27.17	45.99

Conclusions

Contents

Conclusions

Summary

- Translation at sentence level misses important information for coherence at **document level**
- **Data-based** and architecture-based document-level NMT systems are being studied
- We analyse systems which include information on **coreference** chains and the **topic** of a document
- We obtained **promising** results for coreferents in a document

Conclusions

Analysis & Future Work

- The improvement on the translation depends dramatically on the quality of the **coreference system**

- Current **heuristics** for annotation miss cases

Conclusions

Analysis & Future Work

- The improvement on the translation depends dramatically on the quality of the **coreference system** $\Rightarrow$ more languages and systems
- Current **heuristics** for annotation miss cases $\Rightarrow$ set of heuristics enlarged

Conclusions

Analysis & Future Work

- The improvement on the translation depends dramatically on the quality of the **coreference system** $\Rightarrow$ more languages and systems
- Current **heuristics** for annotation miss cases $\Rightarrow$ set of heuristics enlarged
- Including coreferences as alignments between consecutive source sentences and keep them in a caché (architecture-based) $\sum$ **2 previous talks!**
- Building a multilingual (currently 6 languages) parallel **corpus** of Wikipedia bios

Addendum

Corpus of Biographies

- Physicists, Linguists & Actors in Wikipedia
  (50% man, 50% woman)
- English, German, Spanish, Catalan, French & Arabic

Addendum

Corpus of Biographies

- Physicists, Linguists & Actors in Wikipedia (50% man, 50% woman)
- English, German, Spanish, Catalan, French & Arabic

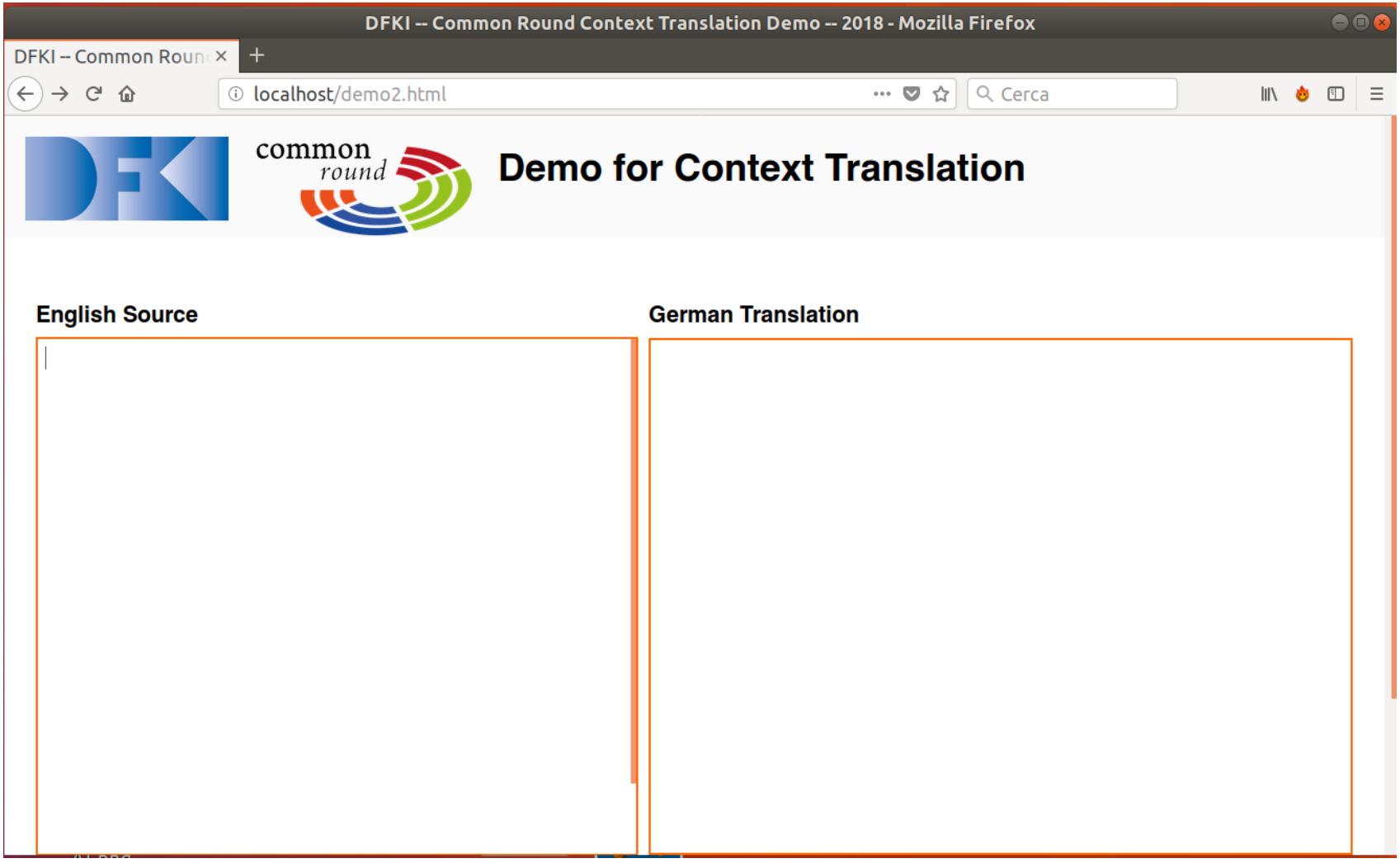
WIKITAILOR

Extraction of multilingual corpora in any domain from Wikipedia

(`https://github.com/cristinae/WikiTailor`)

Demo

Document-Level Translation



Thanks! And...

wait!

Neural Machine Translation with Context & Document Information

Cristina España-Bonet
Universität des Saarlandes (UdS)
Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI)

Guangdong University of Foreign Studies, China

23rd October 2018